REPRINT

Original research article

From observations of individual behaviour to social representations of personality: Developmental pathways, attribution biases, and limitations of questionnaire methods

Jana Uher [*1], Christina S. Werner[2], Karlijn Gosselt[3]

1 Comparative Differential and Personality Psychology, Freie Universität Berlin, Germany
2 Psychological Methods, Evaluation and Statistics, University of Zurich, Switzerland
3 Behavioural Biology, Utrecht University, The Netherlands

*Corresponding author:
Jana Uher
Comparative Differential and Personality Psychology
Freie Universität Berlin
Habelschwerdter Allee 45
14195 Berlin, Germany
mail: uher@primate-personality.net
fon: +49-(0)30-838 55 600
fax: +49-(0)30-838 55 789

Highlights

- New philosophy-of-science paradigm to systematically study both individual-specific behaviours and pertinent representations.
- A non-lexical taxonomic approach used to generate emic personality constructs.
- Over 3 years and 6 waves, 104 crab-eating macaques macaques (Macaca fascicularis) and 99 human observers, experts and novices, studied.
- Attribution biases reflecting socio-cultural stereotypes about age, sex, and social rank demonstrated in personality ratings.
- Important implications for methodology and research methods, in particular limitations of questionnaire methods, illuminated.

Abstract

Socio-cognitive abilities to recognise and to represent individual-specificity—even in some nonhuman species—are central to human life. Using a novel philosophy-of-science paradigm, we explored these abilities over 3 years in 6 waves by investigating individual-specific behaviours of 104 crab-eating macaques (*Macaca fascicularis*) and the representations that 99 human observers—experts and novices—developed of them. By applying the non-lexical Behavioural Repertoire x Environmental Situations Approach, we generated 18 macaque-specific personality constructs. They were operationalised with behavioural measures to study the macaques and with two rating formats to study the observers' representations. Analyses of reliability, cross-method coherence, taxonomic structures, associations with demographic factors, and 12-24-month stabilities highlighted essential differences between individual-specific behaviours and pertinent representations, explored developmental pathways of representations, and illuminated attribution biases and limitations of questionnaire methods.

Key words:
age differences; anthropomorphic bias; attribution bias; Behavioural Repertoire x Environmental Situations Approach; Macaque Personality Inventory for captive populations (MPIc); lexical approach; personality assessment; sex differences – gender differences; social representations; social status

1. Theoretical Background

Knowledge of people in general and of the ways in which individuals differ from one another plays a central role in human life. Person-related information is so important that gossiping about who-is-doing-what-with-whom makes up about two thirds of conversation time (Dunbar, 1996) and occurs in all human cultures (Brown, 1991). Everyday knowledge of individuals is mentally represented in constructs that people develop to describe, to integrate, and to explain their experiences with their social worlds (Kelly, 1955; Valsiner, 2012). By communicating their experiences and personal constructs and by negotiating shared meanings, people create socio-culturally shared ideas, values, and beliefs that are represented in social constructs—that is, they develop social representations (Jovchelovitch, 2007; Moscovici, 1984). Over time, representations of those individual differences that are perceived as most salient and that are considered to be socially relevant in particular communities become encoded in their everyday languages (John, Angleitner, & Ostendorf, 1988).

Constructs of personality differences and their lexical encodings are essential socio-cognitive tools that people intuitively use to quickly form impressions of others (cf. Asch, 1946) based on category systems that have proven to be socially significant within their communities (Goldberg, 1981). Personality constructs are like glasses through which humans peer into their social worlds. They allow people to gain some cognitive control of social interactions with other individuals in their communities—but also with strangers (cf. McAdams, 1994). Dealing with strangers is so pervasive in the everyday lives of present-day humans in large societies that, at first sight, this remarkable ability does not appear to be special in any way. But, in fact, it may be uniquely human (Blaffer-Hrdy, 2009), at least when only mammalian species are considered. In other (non-domesticated) mammals, encounters with strangers regularly result in automatic attack—even in our closest living relatives, the nonhuman primates. “Compared with our nearest ape relations, humans are more adept at forestalling outright mayhem. Our first impulse usually is to get along” (p. 3).

It has therefore been argued that the abilities to recognise and to mentally and socially represent abstract ideas of how individuals generally behave and differ from one another first enabled our human ancestors to deal with unknown others in peaceful ways (the *personality-constructs-promote-peaceful-anonymous-contacts hypothesis*; Uher, 2013). These abilities could have been essential prerequisites for peaceful traffic, exchange, and trade among different socio-cultural communities—behaviours of enormous importance in human evolution. Given this and given our current state of knowledge of other species’ pertinent abilities (see below), the mental construction and social representation of personality differences in and of themselves seem to be uniquely human as well (Uher, 2013).

Most personality constructs that people develop refer to human conspecific individuals. But they are not restricted to them. Humans also recognise and represent individual differences of some other species—and have done so for several tens of thousands of years already. Evidence comes from an impressive 40-year breeding experiment aimed at replicating processes of animal domestication with farm foxes. In these canids, strong selective breeding for a specific behavioural pattern called tamability (i.e., low fearfulness of and low aggressiveness to humans) over just 30-35 generations resulted in a host of changes in genes, morphology, physiology, and behaviour (including non-selected, intra-specific social behaviour) in which the present-day’s domesticated species differ markedly from their wild relatives. These exciting results suggest that the key factors of artificial selections that humans have imposed on some species during domestication consisted of individual behaviours rather than size or reproductive capacity (Belyaev, 1969; Trut, 1999).

Animal domestication—one of the most important developments in recent human history (Diamond, 2002)—thus presupposed that our human ancestors were able to recognise, mentally construct, and socially represent meaningful individual differences in the behaviour of some nonhuman species. Palaeolithic dog fossils in Europe dated back to 30,000 years ago (Germonpré, Sablin, Stevens, et al., 2009) suggest that humans had to

have already developed these abilities by that time. Pertinent representations and semiotic symbols referring to human individuals had very likely already been developed before (Uher, 2013).

The present article explores these remarkable human abilities using a novel research paradigm. This paradigm provides an elaborated philosophy-of-science framework for personality psychology (Uher, 2013). It embarks on a new research strategy by scrutinising these socio-cognitive abilities from meta-theoretical viewpoints and by clearly differentiating the different kinds of phenomena that are constructed as personality. This highlights important implications for research methodology and investigatory methods that are still not well considered in personality psychology.

1.1 The new research paradigm: Meta-theoretical foundations

Elementary to the new research paradigm are meta-theoretical definitions of the phenomena under study, in particular of behaviour. In psychology, definitions of behaviour are rarely discussed and the few proposed are only operational or nominal (e.g., Furr, 2009). The meta-theoretical definition of behaviour as "external activities or externalisations of living organisms that are functionally mediated by the environment (Millikan, 1993) in the present" (Uher, 2013) emphasises that behaviour is inherently bound to the present—and thus requires realtime measurement. Externality differentiates behavioural phenomena from psychological phenomena, which are also bound to the present, but are phenomena entirely internal to the individual. Psychological phenomena can be directly perceived only in oneself through introspection (Wundt, 1904), but not in other individuals (Locke, 1689; Toomela, 2008, 2011). In other individuals, people can directly perceive only behaviours and outer appearances. Psychological phenomena of others, by contrast, can be inferred only from their externalisations, and in particular from behaviour, including parts of language (Uher, in prep. a). But these externalisations may not reflect psychological phenomena, their qualities, or their structures unequivocally or accurately (Cervone, Shadel, & Jencius, 2001; Lewin, 1935; Toomela, 2011). The complex interrelations between behavioural and psychological phenomena can be untangled only if these different kinds of phenomena are explored each in its own right and if a priori assumptions about specific interrelations are avoided. This is still rarely done in psychology (Uher, 2013).

Behaviours (and psychological phenomena as well) are not only bound to the present; they are also dynamic and highly fluctuating. This substantially hinders the recognition of patterns that are specific to individuals because, given these fluctuations, individual patterns can be only probabilistic. But behavioural (time-relative) probabilities that characterise all individuals of a particular population or species are not individual-specific. The probabilities must differ between individuals in stable ways across time periods that are longer than those in which the probabilities were first ascertained. Consequently, individual-specificity refers to patterns that are probabilistic, differential, and temporal (Uher, 2011a). Such patterns cannot be directly perceived. Their recognition requires repeated perceptions of events in many individuals over time and the mental abilities to perceive time and to memorise, to abstract, to (re-)construct, and to represent the experiences made (Uher, 2013).

This complex constellation of abilities seems to be uniquely human. Individualised, non-kin-based dyadic relationships in some nonhuman species suggest that their individual members are able to mentally construct and represent behavioural regularities of *specific* individuals in their social worlds. But they may not be able to generate and mentally represent abstract ideas of how individuals *generally* behave and differ from one another over time, as humans are able to. Without such socio-cognitive category systems, nonhuman individuals cannot quickly form impressions of other individuals, and interactions with strangers are highly unpredictable. This may contribute to the enormous social tension and frequent aggression that we can typically observe in encounters between strangers in nonhuman (non-domesticated) mammals. The abilities of humans to develop social representations of personality differences and pertinent lexical symbols—which substantially facilitate exchange between individually constructing minds and thus the propagation of

socially shared ideas (Lahlou, 1996)—may therefore have played important roles in human evolution (see above; Uher, 2013).

The mental and social processes that are involved in the recognition and construction of individual-specificity entail that personality constructs do not refer only to phenomena that are perceivable in other individuals (i.e., to behaviours including parts of language and outer appearances). Rather, they typically also involve interpretations, appraisals, and explanations of possible causes and consequences of the recognised patterns—particularly in terms of psychological phenomena and environmental factors. These ideas, values, and beliefs are intrinsically embedded into the socio-cultural contexts of particular communities (Wagner, Farr, Jovchelovitch, et al., 1999). The pertinent everyday language terms are therefore loaded with implicit meanings that likely vary socio-culturally (Neuman, Turney, & Cohen, 2012). Consequently, representations of personality differences are far more than mere reflections of individual-specific behaviours and outer appearances. They are constructions of new realities—social realities—that are essentially different kinds of phenomena.

The new research paradigm (Uher, 2013) considers this and explicitly differentiates *a)* behavioural phenomena and other ecto-phenotypical phenomena such as outer appearance, *b)* internal, especially psychological phenomena, *c)* external environmental phenomena, in particular those that are functionally mediating behaviours (see above) and that are defined as environmental situations (Uher, in prep. a), and *d)* pertinent representations that people develop of all of this. The clear differentiation of these different kinds of phenomena and the meta-theoretical analyses of their different theoretical natures point to important implications for research methodology and investigatory methods.

Specifically, behaviours (and outer appearances) and environmental situations are exterospectively accessible. In these phenomena, the demarcation of entities and their encodations as data can therefore be explicitly defined. The particular elements of the set *B* of behaviours, (the set *O* of outer appearances), the set *S* of environmental situations in which these occurred, the set *T* of occasions and spans of time, and the set *I* of individuals that are considered in quantifications of individual-specificity are thus explicitly known. If realtime measurements are accumulated over time using the concept of time-relative probabilities (for details, see Uher, 2013), this allows for ratio-scaled and—in set-theoretic regards—*objective quantifications* of individual-specificity (Uher, in prep. b). A comprehensive empirical application of these concepts to individual-specific behaviours based on 146 contextualised behavioural variables can be found in Uher, Addessi, and Visalberghi (2013).

This is not possible for the pertinent representations, however, because these are inherently subjective and intersubjective phenomena respectively (Jovchelovitch, 2007; Moscovici, 1984). Because they are not exterospectively accessible, the demarcation of entities in these phenomena cannot be explicitly defined. The processes of perception, memorisation, abstraction, and mental and social construction that are involved in the development of representations of individual-specificity entail that the particular elements of the sets *B,(O), S, T,* and *I* that people implicitly consider in their constructs of personality cannot be traced anymore. They are unknown as are the elements of the sets of interpretations, appraisals, and explanations that personality constructs inherently comprise (for detailed discussions, see Uher, 2013).

Judgements of personality require people to directly quantify individual-specificity. But individual-specificity is not a phenomenon of the present because it refers to probabilistic and temporal patterns (see above) that inherently involve events of the past. Consequently, individual-specificity cannot be directly perceived at a present moment—and thus cannot be directly quantified. To fulfil the requirements of personality judgements, people must necessarily rely on their pertinent representations. Thus, when judging a particular individual, we do not know which particular elements of the above-mentioned sets of elements judges implicitly consider and how these elements are demarcated and converted into data (cf. Rosenbaum & Valsiner, 2011). It follows that judgements of individual-specificity can

provide—in set-theoretic regards—only *subjective quantifications* that are at best ordinal-scaled (Uher, in prep. b, 2013).

1.2 A non-lexical emic approach for taxonomic investigations

These insights have important implications for taxonomic research, which has been based almost exclusively on judgements and in many cases on lexical approaches so far. They highlight that important taxonomic models of human personality differences capture representations that people have developed of individual differences and therefore inherently comprise ideas, values, and beliefs that likely vary socio-culturally (see above). But there is still little scientific description of the individual differences that particular socio-cultural communities perceive to be salient and consider to be socially relevant and of the factors and processes that shape the development of pertinent everyday language terms (John et al., 1988). Personality psychologists still do not know how well lexically derived models represent perceivable individual differences, the ways in which they may reflect different perceptions, interpretations, and appraisals; and how perceivable individual differences actually vary within and between different socio-cultural and language communities (Block, 2010; Uher, 2013).

For comprehensive taxonomic investigations of and systematic comparisons between individual-specific behaviours and people's pertinent representations, a new methodological approach has been developed that is grounded in the philosophy-of-science framework of the new paradigm—the Behavioural Repertoire x Environmental Situations Approach (BR_xES-Approach; Uher, 2008a, b, 2011a, b). This approach is a non-lexical emic/bottom-up approach (i.e., a manifest system approach; see Uher, under review) that allows researchers to systematically generate personality constructs from within the known behavioural repertoire of a population—both human and nonhuman. It breaks the generation of constructs of individual-specificity down to observational concepts of behaviours and of environmental situations that scientists have already described for the *average* individual of a study population (see Section 2.3). These hypothetically generated constructs serve to systematically guide the researchers' selections of what to study to allow for comprehensive taxonomic investigations. On the one hand, these constructs can be used to taxonomise individual-specific behaviours of a given population (e.g., of a socio-cultural community or a species); for this purpose they are operationalised in behavioural measurements. On the other hand, BR_xES-Approach-generated constructs can be used to study the representations that a specified group of people have developed of the individual-specific behaviours of a given (human or nonhuman) population. For this purpose, the constructs can be operationalised, for example, by rating items that describe their content in the particular language of the raters under study.

Like any research, the BR_xES-Approach necessarily relies on human language. But in contrast to lexical approaches, its rationale of selection is not guided by the everyday language terms in which particular socio-cultural communities have encoded their pertinent representations. Instead, it is based on scientific terms, descriptions, and categorisations of the known behavioural repertoire of the population under study. Construct labels are therefore less colloquial than those derived from everyday languages. Moreover, representations typically comprise—accurately or not—causal assumptions (see above). Perhaps for this reason, lexically derived constructs are frequently—but erroneously—attributed a causal status (Bock, 2000; Komatsu, 2012; Lamiell, 2003; Mischel & Shoda, 1994; Uher, 2013). Following the clear differentiation of behaviours from both causally related internal and external phenomena and from pertinent representations, the BR_xES-Approach generates descriptive constructs that are not attributed a causal status (Uher, 2011a). This allows for the generation of comprehensive structural-descriptive knowledge of individual-specific behaviours that can systematically guide and meaningfully complement research on causally related phenomena, particularly psychological ones.

1.3 Essential differences between research on individual-specific behaviours and research on pertinent representations

The philosophy-of-science perspective of the new paradigm also highlighted differences in the properties of the data that can be obtained from these different kinds of phenomena. Specifically, individual-specific behaviours may not be as consistent as people's judgements of them. Behavioural data need not fulfil the psychometric standards established for judgement-based research (Uher et al., 2013). This is a reflection of the well-known facts that individual-specificity in behaviour emerges at the fine-grained levels of individually distinct yet stable behavioural profiles across situations (e.g., Mischel, 1977) and across different functionally related behaviours within a situation (e.g., Asendorpf, 1988). These patterns entail that, on the sample level, the cross-situational consistency of individual behaviour (Mischel, 1968) and the consistency between functionally related behaviours can be only moderate (Asendorpf, 1988). These phenomena have also been shown in nonhuman primates (Stevenson-Hinde, Stillwell-Barnes, & Zunz, 1980; Uher, 2011a, 2011b, Uher et al., 2008; Uher et al., 2013).

The patterns of (in)consistency in behavioural phenomena are not readily apparent in judgement-based research. In their experiences with their worlds, people strive to detect recurrent patterns that may enable the predictions of events (Kelly, 1955) while they are facing the uncertainty of the future (the so-called ecological necessity of abstraction; Valsiner, 2012). They may therefore develop somewhat coherent representations that are consistent with the logic of the human mind and with socio-cultural belief systems—yet not necessarily with the phenomena that are being represented (Daston & Galison, 2007; Uher, 2013). The impact of mental and social processes on the structure of judgement-based data of individual differences has received only little consideration so far (Diriwächter, Valsiner, & Sauck, 2005; Rosenbaum & Valsiner, 2011; Schwarz, 1999; Uher et al., 2013).

Additionally, judgement scales of personality inventories and lexically derived constructs are developed by selecting only those variables that yield empirical data structures with high internal consistency in the target population—and that thus measure redundancies (Block, 2010). In everyday language, redundancies can be easily created at low cost (cf. Lahlou, 1996) and are therefore comparably widespread. But in behaviour, redundancies may be rare. They may even be constrained because they are too costly in ecological and evolutionary regards.

The BR_xES-Approach considers these peculiarities of behavioural phenomena and allows researchers to employ a two-step procedure to explore taxonomic structures of individual-specific behaviours. In the first step, the behavioural data are aggregated on the level of the generated constructs (see Section 2.3). This reduction is based on the scientifically established functions of the studied behaviours—regardless of potentially low internal consistencies between their measurements. In the second step, these theoretically derived construct measures are statistically analysed for taxonomic structures. The first step corresponds to the processes of mental abstraction and construction on the part of human observers, but is, in contrast to them, made explicit and based on scientific knowledge. It can thus be traced to the specific behaviours and situations that are considered in these measures (see above, Uher, 2013).

The BR_xES-Approach has already been successfully applied to investigate individual-specific behaviours of capuchin monkeys (Uher et al., 2013) and of great apes (Uher, Asendorpf, & Call, 2008) and to investigate the representations that human observers (i.e., keepers) have developed of great ape individuals (Uher, 2011b; Uher & Asendorpf, 2008). To investigate both kinds of phenomena, it was also applied in the present study.

1.4 The present research

This study explored the human abilities to quickly form impressions of nonhuman individuals. Because domesticated animals have been selected and bred for physical and behavioural properties and performances that humans can easily perceive and thus well represent and because knowledge of these animals is widespread in the public, we studied

crab-eating macaques (*Macaca fascicularis*), a nonhuman primate species endemic to Southeast Asia, in a sample that has been kept in The Netherlands, Europe. Crab-eating macaques, also called long-tailed macaques, cynomolgous, or Java monkeys, live in large groups with one or several males, many females, and their offspring. Their social structure is characterised by female matrilines, pronounced dominance hierarchies, and one alpha male leading the group (Angst, 1975). Their sex dimorphism is pronounced; females have about 69% of average male weight (McDonald, 2001). These monkeys are opportunistic omnivores; they prey on vertebrates, invertebrates, and eggs. They wash food (Visalberghi & Fragaszi, 1990), swim under water, and use stone tools to crack open oysters, bivalves, crustaceans, and nuts (Gumert, Kluck, & Malaivijitnond, 2009). To our knowledge, comprehensive taxonomic investigations of the personality differences of this species are still missing.

Nonhuman primates are among the most interesting species to study for individual-specific behaviours because of their complex behavioural and social systems and their gradient of phylogenetic relatedness to humans—the primate species *Homo sapiens*. The nonhuman primates' many striking similarities with us regularly prompt people to attribute anthropomorphic characteristics to them. But they also show many dissimilarities with human primates by which biased attributions become particularly apparent.

To explore how quickly and accurately human observers may develop representations of individual crab-eating macaques, we investigated persons who were previously inexperienced with this species (i.e., novices) and persons who were well experienced with both this species and the particular individuals under study (i.e., experts). Using BR_xES-Approach-generated constructs, we studied the macaques' individual-specific behaviours using ethological observations and the observers' pertinent representations using personality judgements. We gave the novices five days of intense observation of just five target macaques kept in large social groups and then studied the representations that they developed from these limited observational experiences and those that the experts had developed based on their much broader experiences. We hypothesised that this brief time and the limited sample of individually known macaques would cause substantial differences in the novices' agreement both with one another and with the experts. We used two formats of rating items that differed in their degree of abstraction from perceivable events to systematically explore how the observers represented the macaques' individual-specific behaviours, how they may have developed their representations, and to illuminate possible attribution biases.

2. Methods

2.1 Macaque individuals

We investigated 104 crab-eating macaques (*Macaca fascicularis*) at the Ethology Station of the Behavioural Biology Department, Utrecht University, The Netherlands. The 69 females and 35 males were 1 to 33 years old (*Mdn* = 6.3; *M* = 8.1; *SD* = 6.6). The sample comprised 19 male and 24 female subadults (1-5 years), and 16 male and 45 female adults. The macaques lived in three groups of 39-42 (R-group), 34 (S-group), and 24-27 (T-group) individuals in spacious indoor and outdoor enclosures. They were always treated in accordance with the Guidelines for the Treatment of Animals in Behavioural Research and Teaching (2006) and received their complete daily diets and permanent access to fresh water.

2.2 Human observers: Experts and novices

Two groups of human observers that differed in their levels of experience participated in this study. The *experts* were eight supervisors (all women) who were working at the Ethology Station. They had known the particular macaque individuals for several months up to several years. The *novices* were 91 undergraduate students (35 men, 56 women) who participated in ethology courses conducted at the Ethology Station. The novices were

previously inexperienced with this particular species, with the particular macaque individuals in the study, and with ethological observations.

2.3 Non-lexical generation of emic working constructs of personality differences

We applied the non-lexical BR_xES-Approach (see Section 1.2, Uher, 2008a, b, 2011a, b) to systematically generate emic personality constructs that have high ecological validity for crab-eating macaques. First, we conducted a systematic review of 23 publications about the behavioural repertoire of captive and wild crab-eating macaques available at the start of our study (all references are listed in the Appendix). From these publications, we compiled a large table with all major behavioural categories (listed in one column) together with the categories of environmental situations in which these behaviours are described as commonly occurring (listed in a second column). Each row of the table thus represents a unit of a particular behavioural category and a particular associated situational category as described in a given publication; this is called a *behaviour$_x$situation-unit* in the BR_xES-Approach. The primary compilation of categories was designed to be overinclusive, repeatedly listing the same behaviours and situations. Then we reorganised the table such that categories describing the same or functionally similar behaviours were grouped together. We also organised the associated situational categories within the given behavioural categories. Finally, the *behaviour$_x$situation-units* were organised hierarchically according to the degrees of abstraction with which they describe the behaviours and situations.

Using *behaviour$_x$situation-units* on moderate levels of abstraction that reflect relatively homogeneous subsets of still identifiable concrete behaviours and situations, we generated personality constructs by *hypothetically* assuming individual-specific patterns in the particular contextualised behaviours described. The thus-generated constructs (listed in a third table column) are therefore called *working constructs*. They serve methodological purposes to systematically guide the researchers' decisions of what to study, but they do not a priori imply empirical usefulness. Given the over-inclusiveness of the compilation, the same working constructs were generated repeatedly in different parts of the category system. This is the essential prerequisite of the BR_xES-Approach that enables researchers to systematically generate non-lexical emic personality constructs by considering the entire known behavioural repertoire of the population studied. In a second identical table, we then sorted the rows by the generated constructs and eliminated redundant enumerations of the same constructs to obtain a comprehensive overview of all generated constructs and the major behavioural and situational categories in which they describe individual-specificity. This construct generation process is described in more detail in Uher et al., (2013).

According to the emic/bottom-up reasoning of the BR_xES-Approach, all working constructs were constructed a priori to be unipolar in describing individual-specific behaviours of the *same* function (e.g., low to high Playfulness). Whether several working constructs can be constructed as representing opposite poles of a few more abstract taxonomic constructs that describe behavioural patterns of *distinct* functions (i.e., that are bipolar) is left to empirical investigation in each study population. For the present study on captive Java macaques, this procedure of the BR_xES-Approach yielded 18 non-lexical emic personality constructs. Working constructs describing behaviours and situations that occur only in the wild, such as territoriality and travelling, could not be considered. We also had to exclude the construct "Food orientation" because the monkeys were fed outside observation hours.

2.4 Multi-method operationalisations of working constructs

The generated working constructs were operationalised systematically with different methods. To study the macaques' individual-specific behaviours, behavioural measurements were obtained in ethological observations for most constructs. To study the observers' representations, all 18 constructs were operationalised as trait-adjective and behaviour-descriptive verb items for observer judgements on standardised scales.

2.4.1 Ethological observations (EO)

In systematic ethological observations, data were collected on 34 social and non-social behaviours that regularly occur in the daily settings of captive species-typical social groups. All behaviours were described and defined in a comprehensive ethogram based on the established behaviour-scientific knowledge of crab-eating macaques (see Section 2.3; Appendix). A fixed observation schedule ensured that all target macaques were observed to the same extent. Within each Study wave (see Section 2.5), each target macaque was continuously observed in 17 focal individual time samples of 5 min each that were distributed evenly over five consecutive observation days; overall, 85 min per Study wave. The behavioural records comprised frequencies and durations; for the latter, one-zero records in 15-sec time intervals were used to estimate the intervals that included any amount of time spent in the respective behaviours (Altmann, 1974). Laboratory-based experiments for non-invasive behavioural testing of individuals that would have enabled the collection of ethological data for all BR_xES-Approach-generated constructs were not possible. In total, we obtained one to six behavioural measurements in $N = 101$ individuals for 11 of the 18 working constructs.

2.4.2 Rating instruments: The Macaque Personality Inventory for captive populations (MPIc)—Trait-adjective items (TA) and Behaviour-descriptive verb items (BV)

The items were phrased in the observers' native language (Dutch). The number of items had to be minimised because the observers judged many (up to 37) macaque individuals. Moreover, the sample sizes of macaques that are sufficiently well known as individuals are extremely limited so that variables-to-cases ratios are generally compromised. In both item formats, judgements could be indicated on a 5-point frequency scale from (1) *hardly ever*, (2) *rarely*, (3) *sometimes*, (4) *often,* to (5) *almost always*.

Trait-adjectives (MPIc–TA). Each working construct was operationalised with one specific trait-adjective item that best described its content in the raters' everyday language, yet without explicitly stating any specific behaviours or situations. For example, Curiousness was operationalised with "[Individual's name] is curious". Trait-adjectives are abstract words referring to representations that are distant from immediate perceptions of behavioural and situational events (see Section 1.). They therefore involve more complex processes of mental (re-)construction on the part of the raters than behaviour-descriptive verbs. In total, we constructed 18 trait-adjective items; none of them was reversed in meaning to avoid negations.

Behaviour-descriptive verbs (MPIc–BV). Each working construct was operationalised with one to three verb-based sentences describing specific behaviours and situations that were directly based on the categories used to generate and to define the constructs (see Section 2.3). Behaviour-descriptive verb items specify concrete observable behavioural and situational events and therefore involve less complex construction and inference processes on the part of the raters than the abstract trait-adjectives. For example, Curiousness was operationalised with "[Name] intensely inspects new objects from close-by and/or touches them". Most working constructs were operationalised with two items, seven constructs with one item, and one with three items. Overall, 30 behaviour-descriptive verb items were constructed, five of which could be reversed in meaning without using negations in the item wording.

2.5 Procedures

2.5.1 Ethological observations

In the Ethology courses, the novices first learned about the theoretical and methodical foundations of ethological observation. Then they received practical training, and finally, they conducted their own ethological observations of the macaque individuals housed at the station. The novices worked together in pairs. Each pair jointly observed five target individuals from the same macaque group. Up to ten student pairs worked in parallel, thus

observing up to 50 macaques within one Study wave. The observations of each macaque group were supervised by two experts. They first trained the novice pairs to reliably identify their five target individuals and to systematically observe their behaviours using a comprehensive predetermined ethogram. Then the novices collected behavioural data on five consecutive days of intense and systematic ethological observation (see Section 2.4.1).

2.5.2 Personality judgements

Each novice pair and each expert pair provided judgements on the macaque individuals that they had observed. So that the novices would not focus on individual differences during their ethological observations, we asked the novices to provide personality ratings only *after* they had completed their ethological observations and when they no longer had access to the macaques. Observer judgements were collected via the Internet portal *www.primate-personality.net.* On a leaflet, we informed them about the rating procedure and the online access to the inventories. We explicitly cautioned all observers not to discuss their ratings with the other raters. Because all observers repeatedly provided judgements on the same items for different macaque individuals (the novices for 5 macaques, the supervisors for up to 37 macaques), we presented the inventories in an interactive user interface (programmed by JU) that allowed for the personalised online-presentation of items for each rater. It also inserted the name of the particular target macaque into the wording of each single item to help the raters focus on the particular individual being judged. All 18 trait-adjective items and all 30 behaviour-descriptive verb items were presented together in a fixed randomised order in chunks of five items per web page to avoid cross-checking between responses to items of related content. A similar computerised item presentation had already been successfully applied for keeper judgements of great ape individuals (Uher, 2011b; Uher & Asendorpf, 2008). All raters assessed their target macaques in a predetermined randomised order. Because each pair of novices pair and of experts judged the same set of individuals, one rater per pair provided judgements in alphabetical order of the macaques' names, the other one in inverse order to avoid effects of familiarisation with the inventories on the ratings of single macaques.

2.6 Study waves

The study spanned three consecutive years of systematic multi-method data collection. In each year, we obtained data from ethological observations and personality ratings on both item formats from two ethology courses that were scheduled about three to four weeks apart. Overall, we collected data in six Study waves (t_1-t_6). Given the large number of macaque individuals per group (see Section 2.1), it was not possible to study all macaque groups in all Study waves. We observed the R-group in all six Study waves t_1 to t_6; the S-group in t_1 to t_4; and the T-group in t_5 and t_6. Thus, the number of macaques observed varied slightly between Study waves (72 in t_1; 71 in t_2; 62 in t_3; 61 in t_4; 59 in t_5; 60 in t_6). The experts provided ratings in both Study waves of Study year 1 to allow for analyses of test-retest reliability analyses, and in one Study wave each of the Study years 2 and 3. Overall, up to four raters (i.e., each two novices in t_1 to t_6, and each two experts in t_1, t_2, t_3, and t_6) rated up to 50 macaques per Study wave. The experts also provided ratings on additional macaques that were not included in the behavioural data sets (see Section 2.4.2) and not rated by the novices; these were 52 macaques in Study wave t_2 and 11 in t_3. Thus, the number of macaque individuals rated varied between Study waves (45 macaques in t_1; 97 in t_2; 61 in t_3; 44 in t_4; 49 in t_5; 51 in t_6).

2.7 Data aggregation and data analyses

2.7.1 Technical terminology

The philosophy-of-science research paradigm applied in this study adopts a more technical terminology than is common in contemporary personality research (Uher, 2013). A precise terminology is needed to refer unambiguously to the different phenomena studied (i.e., the macaques' individual-specific behaviours and the representations that the human

observers developed of them) and to the different concepts with and the different levels of aggregation at which these are being described and analysed (e.g., behavioural measurements, behavioural composite construct measures, mean rating scores, see Section 2.7.3). A glossary of the terms relevant to the present analyses is provided at the end of this article.

2.7.2 Levels of aggregation

Ethological observations. Within each Study wave, we first aggregated the behavioural raw data across the five observation days to obtain scores reflecting behavioural probabilities. Because they reflect different types of measure (e.g., frequencies, durations, see Section 2.4.1) and to obtain scores reflecting differential patterns in the probabilities to display these behaviours, we *z*-standardised the aggregate scores within each Study wave. The thus-derived behavioural measurements were then aggregated on the level of the working constructs into behavioural composite measures. To explore whether they reflect in fact individual-specificity (see Section 1.), both the single behavioural measurements and the composite construct measures were analysed for their test-retest reliability between the Study waves within each Study year. Finally, the behavioural composite construct measures were aggregated per Study year.

Observer judgements. For both rating formats, we analysed the raw rating scores for interrater agreement within each Study wave. Thereafter, the rating scores for each macaque on each rating item were aggregated on different levels. First, mean rating scores per Study wave were computed separately for the different rater groups (i.e., experts and novices) and then averaged into mean combined scores of both rater groups. Thereafter, the mean scores of both rater groups were *z*-standardised over individuals within each Study wave and then aggregated into mean scores per Study year. For some analyses, the scores of the behaviour-descriptive verb ratings were further aggregated on the level of the BR_xES-Approach-generated constructs (reversing those of some items to share the same polarity). Finally, the mean scores of all single rating items were statistically summarised into rating factor scores.

2.7.3 Analyses of behavioural data and of rating data

On these different levels of aggregation, we explored our data matrices of *i* individuals by *j* variables from the viewpoint of the variables and studied the *i* individuals' score distributions on the *j* variables using variable-oriented analyses (Stern, 1911). We analysed interrater agreement within each Study wave, and test-retest reliability of the behavioural data and of the rating data between the Study waves of each Study year (Section 3.1). We studied cross-method coherence (Section 3.2) and explored how raters may have formed their adjectival judgements using mediation analysis (Section 3.3). Further, we investigated taxonomic structures of between-individual variations in the behavioural data and in the rating data and analysed the associations between these different data sets. On the level of BR_xES-Approach-generated working constructs, we investigated associations of behavioural composite measures, behaviour-descriptive verb ratings, and trait-adjective ratings with the macaques' demographic factors (Section 3.4). Finally, we studied the 12- and 24-month stabilities of the behavioural composite construct measures and of the rating factor scores (see Section 3.5).

To compute mean correlations and to test correlation scores for differences between methods and rater groups, we always used Fisher's *r*-to-*Z* transformation.

3. Results

3.1 Reliability

3.1.1 Behavioural measurements and behavioural composite construct measures (EO)

The average reliability of the 34 behavioural measurements in terms of their test-retest reliability for the two Study waves of each Study year was r_{tt} = .41 to .46 across all three Study years, ranging for single measurements from r_{tt} = -.07 to .99. The average test-retest

reliability of the behavioural composite construct measures was r_{tt} = .49 to .60, ranging for single measures from r_{tt} = .05 to .93 (Table 1). In two of the three Study years, the test-retest reliability of composite measures was below r_{tt} = .40 for the constructs Impulsiveness (r_{tt} = .13 to .63) and Aggressiveness (r_{tt} = .09 to .45). It was low in all three Study years for the constructs Anxiousness (r_{tt} = .22 to .29) and Dominance (r_{tt} = .05 to .39).

3.1.2 Observer judgements on trait-adjective items (TA) and behaviour-descriptive verb items (BV)

Mean level differences in the judgements provided for the individuals of the three macaque groups were absent; we therefore analysed the ratings of all macaques together.

3.1.2.1 Interrater reliability

Experts and novices showed substantial agreement in their judgements of the *N* = 104 macaque individuals. Across all Study waves and both rating formats, the mean interrater reliability of the average raters (Shrout & Fleiss, 1979) was for experts *ICC*(3,2) = .61; for novices *ICC*(3,2) = .62; and for all four raters from both rater groups *ICC*(3,4) = .72. Across both rater groups, the average interrater reliability of the trait-adjective ratings was *ICC*(3,*k*) = .66, virtually identical to that of the behaviour-descriptive verb ratings of *ICC*(3,*k*) = .64. Interrater agreement did not differ systematically between Study waves (see Table 2). For more direct comparisons between rater groups, we analysed the interrater reliability of the single raters. Across all Study waves and both rating formats, it averaged for experts *ICC*(3,1) = .46; for novices *ICC*(3,1) = .47; and for both rater groups combined *ICC*(3,1) = .42. On trait-adjective items, it was for experts *ICC*(3,1) = .44 and for novices *ICC*(3,1) = .49; for behaviour-descriptive verb items, it was for experts *ICC*(3,1) = .47 and for novices *ICC*(3,1) = .45. The small differences between rater groups and rating formats reflected in these average agreement scores were not significant within each Study wave.

Interrater reliability was low or negative for some rating items in some Study waves and for some rater groups, such as for the behaviour-descriptive verb items “In uncertain situations, [Name] yawns or scratches him/herself” operationalising Arousability, *ICC*(3,1) = -.17 to .32; and “[Name] can occupy him/herself with something for a long time” operationalising Persistence, *ICC*(3,1) = .00 to .22; and for the trait-adjective item “[Name] is cleanly with him/herself” operationalising cleanliness, *ICC*(3,1) = .07 to .28. But no rating item lacked interrater reliability in general in all Study waves or for all rater groups.

3.1.2.2 Test-retest reliability

We analysed the mean ratings at the level of the Study waves for their test-retest reliability within each Study year. Test-retest reliabilities for novice ratings refer to all three Study years; those for expert ratings and for the combined ratings of the two rater groups only to Study year 1 (see Section 2.6). Across the two item formats, the average test-retest correlation for expert ratings (mean of 2 raters) was *r* = .78; for combined ratings (mean of 4 raters), it was *r* = .74; and for novice ratings (mean of 2 raters), it was *r* = .48. Novice ratings showed significantly less test-retest reliability than expert ratings, *t*(47) = 10.093-13.595; *p* = .000, and than combined ratings, *t*(47) = 7.920-15.666; *p* = .000.

Recall that in both Study waves of Study year 1, all expert ratings were provided by the *same* persons, whereas novice ratings were provided by *different* persons in every Study wave. To test for possible effects that this difference may have had on the test-retest reliability of the expert ratings, we computed cross-correlations between the different rater groups (each derived from two raters) within each Study year (i.e., expert t_1-novice t_2, novice t_1-expert t_2, expert t_3-novice t_4, novice t_5-expert t_6). Their average test-retest correlation across the two item formats and all Study years was *r* = .58 (see Table 2). The test-retest correlations of expert ratings were significantly higher than those between different rater groups for both trait-adjective ratings, *t*(17) = 3.098-9.018; *p* = .000-.007, and behaviour-descriptive verb ratings, *t*(29) = 6.957-12.226; *p* = .000, in all cases. These cross-correlations, in turn, tended to be significantly higher than the test-retest correlations

between novice ratings. For trait-adjective ratings, these differences were significant for one expert-novice cross-combination in Study year 1 and one in Study year 2, $t(17) = 2.141$-4.168; $p = .001$-$.047$. For behaviour-descriptive verb ratings, these differences were significant in all three Study years except for one expert-novice cross-combination in Study year 1, $t(29) = 2.636$-7.878; $p = .000$-$.013$.

Across all rater groups and Study years, average test-retest reliability did not systematically differ between trait-adjective items, $r_{TA} = .61$, and behaviour-descriptive verb items, $r_{BV} = .58$ (see Table 2). Four items lacked test-retest reliability in all Study waves and in all rater groups; these were the behaviour-descriptive verb items "In uncertain situations, [Name] yawns or scratches him/herself" operationalising Arousability ($r_{tt} = .34$ for experts; $r_{tt} = .21$ for all raters; $r_{tt} = .09$-$.49$ for novices); "[Name] can occupy him/herself with something for a long time" operationalising Persistence ($r_{tt} = .27$ for experts; $r_{tt} = .01$ for all raters; $r_{tt} = -.12$ to $.24$ for novices); "[Name] cleans his/her skin, fur, and wounds" operationalising Cleanliness ($r_{tt} = .27$ for experts; $r_{tt} = .20$ for all raters; $r_{tt} = -.14$ to $.38$ for novices), and the trait-adjective item "[Name] is cleanly with him/herself" also operationalising Cleanliness ($r_{tt} = -.05$ for experts; $r_{tt} = .04$ for all raters; $r_{tt} = .03$ to $.27$ for novices). Given their lack of test-retest reliability, we excluded these items from subsequent analyses.

3.1.3 Comparison of temporal reliability between the macaques' individual-specific behaviours and the judgements of the different rater groups

The parallel collection of behavioural data and of rating data from different rater groups on the same personality constructs for the same sample of macaque individuals permits direct comparisons of their temporal reliability. We compared the test-retest correlations of expert ratings, novice ratings, combined ratings, and the cross-correlations between the different rater groups on trait-adjective items and behaviour-descriptive verb items with those of the behavioural composite construct measures. This latter set of measures was studied because these composites reflect the highest level of aggregation in the behavioural data. Their level of abstraction from perceivable behavioural events is therefore more comparable to that of the rating data than the level of abstraction of the single behavioural measurements. We tested the differences in test-retest reliability for significance using *t*-tests for independent samples and analysed their magnitude with Cohen's effect size based on pooled standard deviations (Cohen, 1992). We used the scores of Study year 1 in which the largest samples were obtained for all methods.

The test-retest reliability of expert ratings and of the combined ratings was significantly higher than that of the behavioural measures for both trait-adjective ratings, $t(27) = 2.399$-2.722; $p = .011$-$.024$, and behaviour-descriptive verb ratings, $t(39) = 2.512$-3.962; $p = .000$-$.016$. The magnitude of these differences for expert ratings and for the combined ratings of trait-adjective items was $d = 1.04$ and $d = .92$ respectively; and for behaviour-descriptive verb items, it was $d = .86$ and $d = 1.33$, respectively. A further significant yet inverse difference was found for the novices' behaviour-descriptive ratings in Study year 2, $t(39) = -3.422$; $p = .001$. In this exceptional case, the rating data showed substantially less test-retest reliability than the behavioural data ($d = -1.16$). No further differences in temporal reliability were found for either novice ratings or for cross-correlations between rater groups. Thus, only test-retest correlations of scores involving expert ratings in both Study waves of a given Study year were significantly higher than those of the behavioural measures. But this was not the case if expert ratings were involved in just one of the two Study waves compared.

3.2 Validity of observer representations: Cross-method coherence on the level of working constructs

We explored the relationships between the macaques' individual-specific behaviours and the observers' pertinent representations on the level of BR_xES-Approach-generated working constructs in terms of the coherence between their behavioural composite measures and the two judgement-based measures (i.e., trait-adjective ratings and behaviour-descriptive verb ratings). The latter were analysed jointly for both rater groups

using the combined ratings of Study year 1. Across all working constructs studied, coherence between these three methods in terms of Pearson correlations (r) was substantial and differed significantly from zero in one-sample t-tests, $t_{TA–BV}(17) = 7.99$, $p < .001$; $t_{TA–EO}(10) = 6.50$, $p < .001$; $t_{BV–EO}(10) = 10.69$, $p < .001$ (for details, see Table 3). The strength of coherence between methods differed significantly such that, across constructs, the correlations between the two rating methods (mean $r_{TA–BV} = .79$) were significantly higher than those between trait-adjective ratings and behavioural measures (mean $r_{TA–EO} = .45$; $t_{TA-BV – TA-EO}(10) = 3.46$; $p < .001$) and than those between behaviour-descriptive-verb ratings and behavioural measures (mean $r_{BV–EO} = .52$; $t_{TA-BV – BV-EO}(10) = 2.59$, $p < .027$). The latter were not significantly higher than the correlations between trait-adjective ratings and behavioural measures ($t_{TA-EO – BV-EO}(10) = .96$, $p > .360$).

The construct Social orientation to group members showed an interesting pattern of cross-method coherence. The trait-adjective rating ("[Name] is friendly to group members") showed a zero correlation to the corresponding behavioural construct measure composed of the social contact behaviours Groom, Embrace, and the prosocial facial displays Lip-smack and Scalp-lift (see Table 3). This is virtually identical to the results of a methodologically analogous study on keeper judgements of great ape individuals conducted in the German language (Uher & Asendorpf, 2008). In that study, the behaviour-descriptive verb ratings of Social orientation/Friendliness (to group members) were substantially correlated with the corresponding composite behavioural measures ($r = .68$, $p < .01$), but both were uncorrelated with the corresponding trait-adjective ratings ($r = -.06$ to $.00$, n.s.), suggesting that keepers may have based their judgements of "friendly" on low aggression instead of on social contact behaviours (Uher, 2011b).

We therefore further explored the associations between the constructs Social orientation and Aggressiveness (both to group members) across methods. In fact, the trait-adjective ratings of Social orientation ("friendly") were negatively correlated with the Aggressiveness ratings in both formats ($r_{TA(SO) – TA(AG)} = -.54$; $p = .000$ and $r_{TA(SO) – BV(AG)} = -.49$; $p = .000$), but they were uncorrelated with the behavioural Aggressiveness measures ($r_{TA(SO) – EO(AG)} = -.21$; $p = .070$). The behaviour-descriptive verb ratings of Social orientation, in turn, were uncorrelated with all three operationalisations of Aggressiveness ($r_{BV(SO) – TA(AG)/ BV(AG)/ EO(AG)} = .05$ to $.22$; $p = .060$ to $.614$). But conversely, Aggressiveness ratings in both formats were positively associated with the behavioural Social orientation measures ($r_{TA(AG) – EO(SO)} = .37$; $p = .001$; $r_{BV(AG) – EO(SO)} = .34$; $p = .003$). This corresponds to the strong positive correlation that we found between the behavioural composite measures of both constructs ($r_{EO(AG) – EO(SO)} = .72$; $p = .000$).

3.3 Mediation analyses: How observers may have developed representations of the macaques' personality differences

We analysed potential developmental pathways of representations of personality. Specifically, we analysed whether the observers may have developed abstract representations of the macaques (as studied with trait-adjective ratings) rather directly from observations of a broad range of behavioural events (as studied with the behavioural composite construct measures), or whether their more specific representations that referred to only a few indicative behaviours (as studied with behaviour-descriptive verb ratings) may have served as mediators. Partial mediation would be evidenced when, controlling for behaviour-descriptive verb ratings, the behavioural composite construct measures still directly affected the trait-adjective ratings, and complete mediation when they no longer directly affected the trait-adjective ratings.

We estimated and tested this model using multiple regression analyses according to Baron and Kenny (1986) for all 11 working constructs studied with all three methods. The behavioural composite construct measures as predictors were significantly correlated ($p < .001$) with both the trait-adjective ratings as criteria and the behaviour-descriptive verb ratings as potential mediators for eight constructs and for two further constructs when the significance levels were relaxed to $p < .05$. All constructs but Social orientation to group

members (see above) fulfilled these preconditions. For four working constructs, multiple regressions of trait-adjective ratings on behavioural composite construct measures and on behaviour-descriptive verb ratings *(a)* showed a significant impact of the mediator (behaviour-descriptive verb ratings) on the criterion (trait-adjective ratings), and *(b)* rendered the effect of the behavioural construct measures on the trait-adjective ratings non-significant. In these cases, the effects of behavioural construct measures on trait-adjective ratings were fully mediated by behaviour-descriptive verb ratings. For four additional constructs, the behavioural construct measures still directly affected the trait-adjective ratings when controlling for behaviour-descriptive verb ratings, thus fulfilling the criteria of partial mediation (Table 3). These findings mirror previous results on keeper judgements of great ape individuals (Uher & Asendorpf, 2008).

3.4 Taxonomic structures

3.4.1 Intercorrelations between behavioural composite construct measures

Following the two-step reduction strategy of the BR_xES-Approach (see Section 1.2), we studied structural patterns in the behavioural data at the level of working constructs. First, we analysed the internal consistency of the behavioural composite construct measures of Study year 1. In accordance with previous findings and theoretical considerations (see Section 1.1), it was moderate. For those eight construct measures that were composed of two to six behavioural measurements, the average internal consistency of the average measurements was *ICC(*3,*k)* = .627 (range .375 to .827) and of the single measurements it was *ICC(3,1)* = .368 (range .167 to .614). As one would expect, constructs composed of more *k* measurements tended to be more internally consistent, but this effect was not significant (r = .435, p = .282).

We then computed intercorrelations between the behavioural composite construct measures within Study year 1 and Study year 2 in which the same macaque groups (R- and S-groups) were observed (see Section 2.6). The intercorrelational patterns between working constructs showed substantial agreement between these two Study years, thus indicating robust patterns. But there were also some differences. For example, the above-discussed high correlation between behavioural composite measures of Social orientation and Aggressiveness (both to group members; see Section 3.2) was lower in Study year 2, but still significant (r = .32). Behavioural composite measures of Arousability and Social orientation to group members were associated in Study year 1 (r = .48), but not in Study year 2. The same was true for the intercorrelations of the behavioural composite measures of Impulsiveness and Sexual activity, and of Social orientation to group members with both Impulsiveness and Sexual activity (see Table 4).

3.4.2 Exploratory R-factor analyses of personality judgements

We subjected the trait-adjective ratings and the behaviour-descriptive verb ratings on those 44 items that showed both interrater and test-retest reliability to exploratory R-factor analyses. To increase the reliability of the analysed ratings, we used the combined ratings of the two rater groups. We used the first ratings that we obtained for N = 104 macaque individuals; n = 97 of these ratings stemmed from Study year 1, n = 2 from Study year 2, and n = 5 from Study year 3. None of the macaques that were included from the last two Study years were rated again in any other Study year; their inclusion in the structural analysis therefore cannot interfere with analyses of stability between Study years (see Section 2.6).

We applied principal axis factoring with oblique promax rotation, which aims for simple structures and allows for possible intercorrelations at the latent factor level. Based on principal axis factoring with squared multiple correlations as communality estimates, parallel analysis suggested the extraction of four rating factors; results based on principal components analysis were virtually identical. This solution agreed with the graphical elbow in the scree plot. Mean item communality was .70 (range .27 to .91); all but six communalities exceeded .50 (see Table 5). The eigenvalues of the four factors were 11.80, 10.96, 4.51, and 3.68, respectively, corresponding to 27%, 25%, 10%, and 8% in explained item

variance. All four rating factors together explained 70% of the item variance. The rating factors were moderately interrelated; the intercorrelations were $r_{F1\text{-}F2}$ = .12; $r_{F1\text{-}F3}$ = .38; $r_{F1\text{-}F4}$ = .01; $r_{F2\text{-}F3}$ = .04; $r_{F2\text{-}F4}$ = .23; and $r_{F3\text{-}F4}$ = .27, corresponding to a maximum of 14% common variance.

The meaning of the items with dominant loadings allowed for clear interpretations of the four rating factors (see Table 5). The first rating factor labelled *Playful-active-curious* mainly explained items operationalising the BR_xES-Approach-generated working constructs Playfulness, Physical activity, Curiousness, Vigilance, Impulsiveness, and Arousability. The second rating factor labelled *Aggressive-competitive* mainly explained items operationalising Aggressiveness to group members, Competitiveness, Intervening in third-party conflicts, Sexual activity, Dominance, and (inverse) Social orientation to group members. The third rating factor labelled *Prosocial-gregarious* explained items operationalising Gregariousness, Social orientation to group members, and Social orientation to youngsters. The fourth rating factor labelled *Assertive–nonanxious* explained items operationalising Dominance and (inverse) Anxiousness.

For 12 out of 18 working constructs, trait-adjective ratings and behaviour-descriptive verb ratings of the same working construct showed their highest loadings on the same rating factor, thus supporting the above-described evidence that observer judgements in that used different rating formats converged notably across rating formats for most constructs (see Section 3.2). For Anxiousness, Dominance, Impulsiveness, Social orientation to group members, and Social orientation to youngsters, items that were supposed to operationalise the same working construct loaded on different factors. Considering the item content, however, many of these split loadings are meaningful. For example, the trait-adjective item for Impulsiveness “[Name] is impulsive” loaded high (.82) on the *Playful-active-curious* rating factor. The corresponding behaviour-descriptive verb item “[Name] shakes trees, jumps on, or slaps others all of a sudden” did not load on this rating factor at all (.09), but loaded high (.74) on the *Aggressive-competitive* rating factor. The behaviour-descriptive verb item for Social orientation to youngsters that describes the care-taking related aspects of this construct (“[Name] takes care of youngsters by grooming and embracing them”), loaded moderately on the *Prosocial-gregarious* rating factor (.30). But it loaded negatively on the *Playful-active-curious* rating factor (-.26) on which the second behaviour-descriptive verb item for this construct describing the play-related aspects of this construct (“[Name] plays with youngsters”) loaded high (.81). Different kinds of Social orientation to youngsters also seem to be reflected by the corresponding trait-adjective rating item (“[Name] is friendly to youngsters”), which loaded (negatively) highest (-.42) on the *Aggressive-competitive* rating factor, but almost equally high (.41) on the *Playful-active-curious* rating factor.

Operationalisations of the working construct Dominance seemed to be split, however. The trait-adjective item loaded high (.76) on the *Aggressive-competitive* rating factor, whereas both behaviour-descriptive verb items loaded high (-.70 and .61) on the *Assertive–nonanxious* rating factor. The only moderate correlation between these factors (r = .23) suggests that these items may refer to representations of different kinds of dominance (e.g., of dominance as social status versus as assertiveness). The trait-adjective operationalising Social orientation to group members loaded (negatively) moderate to high (-.50) on the *Aggressive-competitive* rating factor, whereas the two pertinent behaviour-descriptive verb items loaded high (.78 and .88) on the *Prosocial-gregarious* rating factor. This reflects the above-analysed patterns of association between the constructs Aggressiveness and Social Orientation to group members. The behaviour-descriptive item describing self-sexual activity as part of the working construct Sexual activity showed multiple low loadings on all rating factors (Table 5).

For subsequent analyses, we estimated factor scores on the four rating factors for all 104 individuals using Thurstone’s (1935) exact regression method (Grice, 2001a, 2001b).

3.4.3 Associations between behavioural composite construct measures and rating factor scores and between their structures

First, we analysed the relations of the behavioural composite measures of the BRxES-Approach-generated working constructs to the four rating factor scores. These associations, analysed with Pearson correlations (*r*), paralleled the factor-analytic loading patterns of the respective rating items in most cases. *Playful-active-curious* rating factor scores were most strongly associated with the behavioural measures of Playfulness (*r* = .81) and Impulsiveness (*r* = .45), and *Aggressive-competitive* rating factor scores with the behavioural measures of Aggressiveness (*r* = .63) and Sexual activity (*r* = .51). *Prosocial-gregarious* rating factor scores were most closely related to the behavioural measures of Gregariousness (*r* = .57) and Playfulness (*r* = .49), and *Assertive-nonanxious* rating factor scores with the behavioural measures of Dominance (*r* = .63) and (inverse) Anxiousness (*r* = -.51; Table 6).

We then compared the associations between working constructs that we found in the behavioural composite construct measures (3.4.1.) with those reflected in the four-factorial rating structure (3.4.2). The structures of these different data sets generally showed considerable coherence (see highlighted cells in Table 6). But we also found some additional associations in the behavioural measures that were not reflected in the ratings. For example, we found correlations between the behavioural composite measures of Playfulness and Anxiousness (*r* = .43 and .55), Arousability and Sexual activity (*r* = .38 and .52), Curiousness and Sexual activity (*r* = .45 and .55), Gregariousness and Anxiousness (*r* = -.25 and -.40), Gregariousness and Dominance (*r* = .36 and .52), in addition to those of Social orientation and Aggressiveness (each to group members) described above.

Conversely, some associations reflected in the rating factor structure (see Table 4) were not found in the behavioural composite construct measures. For example, ratings of Arousability loaded negatively (-.40) on the *Assertive–nonanxious* rating factor; but the behavioural composite measures of Arousability was uncorrelated with those of either Dominance or Anxiousness. Ratings of (allo-)Sexual activity (.75 and .76) and Dominance (.76) loaded high on the *Aggressive-competitive* rating factor, but the corresponding behavioural composite measures were not associated with one another. Ratings of Aggressiveness (.88) and Anxiousness (-.76) loaded high on this rating factor as well, but the behavioural composite measures of these constructs were uncorrelated.

3.4.4 Associations of the macaques' demographic factors with their individual-specific behaviours and how these were represented by the observers

We studied three demographic factors: age, sex, and social rank. The age and sex of all macaques were known from the colony studbooks. Information on social rank was taken from consensual lists of the macaques' linear rank positions that the station's scientific staff maintained and regularly updated based on their observations in research and daily management. These lists were derived from assessments that corresponded to linear Q-sorts of the rank-order of all macaques within each group on dominance. Note that these Q-sorts of social rank were established independently from the dominance ratings obtained from the respective trait-adjective and behaviour-descriptive verb items. These latter measures were provided separately for each macaque using frequency ratings, whereas the Q-sorts were derived from direct comparisons between the macaques. For comparisons across all three groups, which differed in size (see Section 2.1), we first *z*-standardised the social rank data within each group and then pooled the *z*-standardised scores from all three macaque groups to compute a social rank index. To control for the asymmetry in the sex ratio of the sample, which generally comprised more females than males but slightly more males than females in the younger age groups (see Section 2.1), we used multiple regression analyses to explore the relations of the macaques' age, sex, and social rank. We studied the associations of these demographic factors at the level of working constructs using the individuals' scores on the behavioural composite construct measures, the behaviour-descriptive verb ratings, and the trait-adjective ratings of Study year 1 (cf. 3.1.3).

We found substantial cross-method coherences in these associations for many constructs, but also interesting divergences for some constructs. On the one hand, we found associations reflected in the judgements that were not found in the behavioural composite construct measures. For example, in the observers' judgements in both item formats, young age was strongly associated with high Curiousness (*β* = .57-.64; standardised *β* coefficients are provided) and high social rank was associated with high Impulsiveness (*β* = .38-.74) and high Sexual activity (*β* = .42-.43); but such associations were not found for the behavioural measures. Similarly, for the trait-adjective judgements, young age was substantially associated with high Arousability and high Impulsiveness (*β* = .56 and *β* = .52) and high social rank was associated with high Cleanliness (*β* = .26). For the behaviour-descriptive verb judgements, male sex was associated with low Cleanliness (*β* = -.34) and high Sexual activity (*β* = .29). But in all these cases, such associations were not found for the behavioural measures (see Table 7). Conversely, the behavioural measures yielded associations that were not reflected by the observers' judgements. For example, young age was associated with high Anxiousness (*β* = -.32), and male sex with both high Anxiousness (*β* = .21) and high Impulsiveness (*β* = .30). But these associations were not reflected by either rating format.

For some constructs, the patterns of cross-method coherence diverged between item formats. For example, for the behavioural measures, high social rank and high Arousability were positively associated (*β* = .25). This was also reflected by the trait-adjective ratings and was even more pronounced (*β* = .45); but for the pertinent behaviour-descriptive verb ratings, this association was negative (*β* = -.34). By contrast, the finding that high social rank and high Social orientation to group members were positively associated in the behavioural measures (*β* = .36) was also reflected by the behaviour-descriptive verb ratings (*β* = .38), but not by the pertinent trait-adjective ratings (n.s.). For the behavioural measures, male sex and high Social orientation to group members were negatively associated (*β* = -.24). But this was not reflected by the behaviour-descriptive verb ratings (n.s.), and the pertinent trait-adjective ratings even reflected a positive association (*β* = .39; see Table 7 for all standardised *β* coefficients and *p* values).

3.5 Stability across 12 and 24 months

3.5.1 Stability of behavioural composite construct measures

All behavioural composite measures of the BR_xES-Approach-generated working constructs showed significant stabilities across at least one of the two 12-month intervals (mean *r* = .54 and .57 for the first and second 12-month intervals, respectively; ranging from *r* = -.04 to .86). Playfulness was highly stable from Study year 1 to Study year 2 (*r* = .86), but unstable from both Study year 1 and Study year 2 to Study year 3 (both *r*s = -.04). This dramatic decline may be due to the fact that breeding was stopped in those years; therefore, fewer youngsters were available as potential play partners for the individuals studied in Study year 3. But all other behavioural composite construct measures were notably stable even across 24 months (mean *r* = .61; range *r* = -.04 to .85; see Table 8).

3.5.2 Stability of rating factor scores

To study stability across Study years, we computed factor score weights based on the first ratings that we obtained for *N* = 104 individuals (see Section 3.4.1). Using these factor score weights, we estimated factor scores for all subsequent Study years in which rating data for the respective macaques were available (see Section 2.6). This may constitute a conservative estimate of stability at the latent factor level because the rating factor score weights are optimal estimates only for Study year 1 and are influenced by sampling error. We computed longitudinal Pearson correlations (*r*) of these score estimates for each rating factor. Note that these analyses were based on different sample sizes in each Study year (see Section 2.6).

The average 12-month stability of the rating factor scores was *r* = .77 and *r* = .60 for the first and second 12-month intervals, respectively. The average 24-month stability was *r* =

.54. All stabilities were greater than $r = .40$. Stabilities across 12 months were generally the highest from Study year 1 to Study year 2. This may be due to changes in sample size (Study years 1→2, $n = 59$; Study years 2→3, $n = 30$; Study years 1→3, $n = 53$) and in group composition in Study year 3, a change that may also have affected the 24-months stabilities. The most stable scores across all longitudinal intervals studied were found for the *Playful-active-curious* rating factor; the second most stable were those on the *Aggressive-competitive* factor. In addition to the stability of the latent structures underlying these judgements, this may have resulted from the fact that more items had high loadings on these two rating factors than on the two others, thus rendering the score estimates more reliable. The scores on the rating factors of *Prosocial-gregarious* and of *Assertive–nonanxious* showed similar stabilities (see Table 8).

4. Discussion

This study explored the human socio-cognitive abilities to recognise and to represent individual-specificity even in some nonhuman species—abilities that have been essential prerequisites for important developments in recent human evolution, such as for animal domestication (Uher, 2013; cf. Belyaev, 1969; Diamond, 2002; Trut, 1999). We successfully applied a novel philosophy-of-science paradigm for personality psychology to systematically study and compare individual-specific behaviours of crab-eating macaques and the representations that human observers with different levels of experience developed of them. The selection of members of a nonhuman species as those individuals whose individual-specificity is in the focus allowed us to demonstrate essential differences between these different phenomena, to explore the developmental pathways of these representations, and to highlight important implications for methodology and research methods.

4.1 Rapid formation of personality impressions of macaque individuals

In their personality judgements of the macaque individuals, the novices agreed with one another and with the experts equally well as did the experts with one another—in all six Study waves comprising 91 novices. This robust finding is remarkable given that the novices had just five observation days, whereas the experts had already known these macaques for at least several months. The use of ethological methods and of a comprehensive ethogram may have systematically directed the novices' attention to important behaviours of this species, thereby facilitating their acquisition of pertinent knowledge. This finding is also remarkable given that each novice pair focused on just five target macaques. We had expected that knowledge of just a handful of individuals of a previously unfamiliar species would be insufficient to recognise individual-specificity—given that this presupposes knowledge of how individuals of this species generally differ from one another (see Section 1.). At least we had expected agreement between novices to be significantly lower than that between experts who knew all macaques of their study groups. This was not the case. Perhaps, the incidental observation of many "anonymous others" in the target macaques' groups (consisting of 24-42 members) also helped the novices to acquire pertinent knowledge to sufficient degrees.

Despite this, the temporal reliability of judgements provided by the same persons (i.e., experts) was significantly higher than judgements provided by different persons (i.e., between experts and novices and between novices). One explanation may lie in the experts' much greater observational experiences that covered both Study waves compared (and further periods), whereas those of the novices covered just one. Partial overlap in observation periods can also explain why temporal cross-correlations between expert ratings and novice ratings tended to be significantly higher than the temporal correlations between novice ratings, which covered non-overlapping time periods. (The combined ratings of the two rater groups capitalised on the experts' experiences from both Study waves and on aggregation across more raters.)

These findings demonstrate the novices' ability to rapidly develop representations of individual-specific behaviours of crab-eating macaques that were shared with experts on this

species. This supports assumptions and previous indirect (see Section 1.) and direct evidence (e.g., Capitanio, 1999; Pederson, King, & Landau, 2005; Uher, 2011b) that humans are able to apply their sophisticated socio-cognitive abilities to recognise and to mentally represent individual-specific behaviours in socially shared ways to some nonhuman species as well.

However, agreement between persons in and of itself cannot reveal the mental and social processes that are involved in the recognition and construction of individual-specificity nor does it imply the accuracy of these representations in terms of their coherence to the phenomena that are being represented. Present-day humans are born into a world full of complex social knowledge and pertinent semiotic systems. The personal and social representations of human individuals therefore develop in tight dialectical interplays with one another, with the pertinent lexical encodings (cf. Lahlou, 1996, 2001, 2008), and with the individual-specific behaviours (and outer appearances) to which they refer as shown by the pervasive influences that socio-cultural norms and values have on them. These different phenomena cannot be explored and understood independently from one another (Uher, 2013)—at least in studies focussing solely on humans. The present research on a nonhuman primate species that is related to humans yet not endemic to our observers' region of origin opened up interesting opportunities to illuminate these human abilities because these persons could not have relied on complex everyday knowledge of individuals of this particular species.

4.2 Through human personality glasses: How the observers may have developed their representations of the macaques' personality differences

The observers' agreement in their judgements of the macaque individuals is a reflection of their shared socio-cognitive abilities and their necessarily anthropocentric perspective on them. Systematic contrasts to the macaques' behaviours therefore highlighted factors that are important for the recognition, construction, and social representation of individual-specificity.

4.2.1 Salience

One important factor is salience for humans. For example, the constructs Anxiousness, Aggressiveness, Impulsiveness, and Dominance describe behaviours that involve noise, rapid movements (e.g., Scream, Chase, Bite, Tree shake), or striking facial expressions (e.g., Bared teeth) that immediately capture attention. Humans may recognise and represent individual differences in such behaviours more quickly and perhaps even in more pronounced ways than individual differences in behaviours with lower salience. In fact, the trait-adjective ratings of these constructs were amongst the most reliable ones, whereas the corresponding behavioural measures were actually the least or not even temporally reliable. Moreover, these are behaviours of brief occurrence (i.e., frequency or point behaviours) that can be missed out more easily than duration behaviours—in particular in observations of (compared to humans) small and agile individuals in groups of 24 to 42. Conversely, individual differences in arousal behaviours such as Yawn and Scratch, both frequency behaviours of lower salience for humans, were substantially reliable over time, but the corresponding judgements referring to exactly these behaviours showed neither interrater nor test-retest reliability. Individual differences in Self-groom, a non-salient duration behaviour, were also substantially reliable across time, but the observers' pertinent representations were not. These deviations are insightful because the observers (at least the novices) must have consciously perceived the single events—of both salient and non-salient behaviours—that are summarised in the behavioural measures when recording those of their target macaques during observations.

4.2.2 Possible pathways of mental abstraction

A further factor contributing to agreement between observers are their shared human abilities to abstract from perceived behavioural events and to mentally reconstrue individual-

specificity (see Section 1.). Our results on mediation effects suggest (at least for some constructs) that more specific representations of individual-specificity in certain behaviours may have served as intermediate steps in the development of more abstract representations that are frequently encoded with trait-adjectives. For example, observations of events of various aggressive behaviours (e.g., Threat, Chase, Bite) may have been *abstracted bottom-up* into representations of individual-specificity in specific aggressive behaviours that in turn may have facilitated the development of abstract representations of individuals as differing in degrees of "being aggressive". These results mirror previous ones obtained on keeper judgements of great ape individuals (Uher & Asendorpf, 2008). They also match developmental models of impression formation suggesting that, at low levels of experience with a target person, impressions are represented as behavioural exemplars and that from these, with increasing experience, abstract impressions are extracted that can then be retrieved independently for judgement purposes (Park, 1986; Sherman & Klein, 1994).

4.2.3 Socially shared knowledge of human personality differences

To develop representations of personality, however, present-day humans need not individually make complex abstractions from their own experiences. They can also acquire this knowledge from others—through social exchange and from pertinent lexical encodings (cf. Lahlou, 2001). The study of nonhuman individuals provides interesting insights into these alternative pathways of developing representations of personality because there is no reason to assume that humans have developed equally comprehensive systems of knowledge and of pertinent lexical encodings about individuals of other species (Uher, 2011b; except for comparably small lexical repertoires within specific communities, such as those about dogs used in cynology). Hence, to make up for the lack of pertinent knowledge of the nonhuman individuals under study, our observers could have developed knowledge of them based on their comprehensive everyday knowledge of *human* individuals. Social representations and their everyday language terms are important tools of thought and social communication (Neuman, Turney, & Cohen, 2012; Peirce, 1902, CP 4.227; Vygotsky, 1962) that may influence how people perceive, interpret, and appraise individuals and their behaviours. The observers' constructs of human personality—their human personality glasses—could therefore have *guided and shaped top-down* their perceptions of the macaques and their reconstructions and representations of these individuals' personality differences from the very start.

We found interesting pieces of evidence supporting this assumption with regard to some behaviours. For example, the observers' behaviour-descriptive representations of Social orientation were moderately related to the macaques' individual-specific prosocial behaviours, but their pertinent abstract trait-adjectival representations were completely unrelated to these behaviours. Analyses of associations with Aggressiveness illuminated this puzzling finding. In the behavioural data, we found that macaques showing more prosocial behaviours also showed more aggressive behaviours than others. But the observers represented individual differences in these behaviours on behaviour-descriptive levels as unrelated and on abstract trait-adjectival levels as even negatively associated. However, they must have at least implicitly noticed the positive associations reflected in the behavioural data because their Aggressiveness ratings were positively associated with the macaques' prosocial behaviours. But conversely, their Social orientation ratings were completely unrelated to the macaques' aggressive behaviours. These asymmetric relationships indicate that the formation of these representations may have been influenced by socio-culturally shared appraisals of behaviour.

4.2.4 Socio-culturally shared interpretations and appraisals of behaviour

The abstract bipolar representation of Social orientation and Aggressiveness follows the logic of the opposite valence of the trait-adjectives "friendly" and "aggressive" (at least for the observers' socio-cultural communities) but not the structural patterns of the macaques' corresponding behaviours. The implicit meaning of the term "friendly" and the positive

valence attributed to prosocial behaviours obviously precluded the idea that the more "friendly" macaques could also be the more "aggressive" ones. This led to representations that not only failed to reflect the associations found in behaviour, but also to some that reflected even inverse associations. These inaccurate representations could also be an effect of social desirability that people attribute to prosocial behaviours in general and in particular to their interpretation as "friendly" in terms of halo-effects.

These findings also support assumptions that implicit meanings and valences that particular behaviours have for (particular) human observers may interfere top-down with their perceptions and (re)constructions of individual-specificity, thus introducing biases. This top-down interference may also account for the low and highly inconsistent cross-method coherence that we found for the construct Social orientation. By contrast, the high and consistent cross-method coherence of the construct Aggressiveness supports assumptions that the pertinent representations could have been developed from systematic bottom-up abstractions of perceived behavioural events. These findings demonstrate the importance of generating unipolar constructs that describe functionally homogeneous sets of behaviours as conceived in the BR_xES-Approach (see Section 2.3).

4.2.5 Socio-culturally shared assumptions about associations with demographic factors

Socio-culturally shared interpretations, appraisals, and explanations of observable individual behaviour may be particularly pronounced with regard to demographic factors because these are rather directly apparent. The relations between demographic factors and individual-specific behaviours that can be found in nonhuman species may differ from those found in human communities because of differences in their social and behavioural systems. Observer representations that fail to accurately reflect these associations of other species may be due to low salience for human observers—not just of behaviours, but sometimes also of the age, sex, and social rank of nonhuman individuals. But inaccurate representations may also derive from attribution biases—in particular if they reflect pronounced associations not found in the target species. Because "there is a universal tendency among mankind to conceive all beings like themselves" (Hume, 1757/1957, p. xix), such deviations are insightful about the pertinent socio-culturally shared stereotypes of the particular human observers.

Our observers, who all knew at least the sex and age of their target macaques, represented young macaques as being more excitable, more curious, and more impulsive than older ones, males as being less apt to clean and groom themselves and more sexually active than females, and high-ranking macaques as being more impulsive and more sexually active than low-ranking ones. But we found no age, sex, and rank differences in the described behaviours at all. The observers also represented males as being more socially oriented to their conspecifics than females when in fact they were even less so. Conversely, in the behavioural measures, we found that young macaques were more anxious than older ones and males were more anxious than females—but the observers did not represent these (or any other) age and sex differences in these salient behaviours. These biases most likely reflect influences of their socio-culturally shared age- and gender-related stereotypes about *human* individuals—and thus, these are anthropomorphic attributions. These findings provide empirical evidence for previously voiced concerns that relations with demographic factors of nonhuman species found in judgement-based studies (e.g., King, Weiss, & Sisco, 2008) are prone to anthropomorphic biases and may reveal little about the nonhuman individuals under study (Uher et al., 2013).

Interestingly, the observers did not represent the higher impulsiveness of males. Given that impulsive behaviours are salient (see above) and that this finding conforms to socio-cultural gender stereotypes about human males, it may well be possible that the observers did notice and represent the males' higher impulsiveness. They may even have considered this implicitly in their judgements—by comparing males not with all of the macaques in the

sample, but with only other males instead. Thus, they may have adjusted their judgements to a more specific reference population, thereby zeroing out sex differences in the rating data.

This problem and further ones are inherent to judgement-based methods. Their systematic investigation and the development of possible solutions and alternative methods is hindered by the wide-spread but erroneous assumption that questionnaire data could *directly* reflect individual-specific behaviours—and even the psychological *processes* underlying them (Gillespie & Zittaun, 2010; Lamiell, 2003; Michell, 1997; Omi, 2012; Schwarz, 1999; Toomela, 2011; Trendler, 2009; Uher, 2013; Valsiner, 2012; Westen, 1996).

4.3 The new philosophy-of-science paradigm for personality psychology

The new philosophy-of-science paradigm emphasises meta-theoretical differentiations of people's (lexically encoded) representations and of different phenomena that are being represented as personality. Together with analyses of the theoretical natures of these different phenomena, this opened up new possibilities to explore the socio-cognitive abilities that are involved in these representations. This also sheds new light on the questionnaire-based methods and the lexical approaches that are predominantly used for their investigation.

4.3.1 Limitations of questionnaire methods

It is well known that people can flexibly adjust their personality judgements to particular circumstances, such as to specific reference groups (Heine et al., 2002), to different socio-cultural norms that are implicitly associated with the same questionnaire items in different languages (e.g., of the NEO-FFI, Veltkamp et al., 2012), or to experimentally induced goals and motivations (Biesanz & Human, 2010). But these findings reveal neither which particular elements of the sets of behaviours, outer appearances, situations, time, and individuals (*B, O, S, T,* and *I*) people actually consider when they subjectively quantify individual-specificity, nor how they demarcate quantifiable entities in these phenomena and how they convert them into data. It also remains unknown which particular elements of the sets of interpretations, appraisals, and explanations may frame their implicit considerations in a particular judgement (see Section 1.; Uher, 2013).

Fixing the contents to be judged does not solve the problem because the same standardised items of personality inventories are associated with different fields of meaning—both within and even more between persons—as this has already been shown for facets of the Five Factor Model and for items of the pertinent inventories (Arro, 2013; Diriwächter et al., 2005; Rosenbaum & Valsiner, 2011). This offers a further explanation for our finding that judgements provided by the same persons had significantly higher test-retest reliability than judgements provided by different persons: The associated fields of meaning of two persons are likely less diverse than those of four persons. Thus, the temporal reliability of judgements on fixed scales provided by the same persons may also be in part a methodological artefact rather than solely a reflection of the stability of people's actual representations of personality.

It also follows that the particular fields of meaning that the persons under study have in mind when providing personality judgements need not be identical to those that the researchers consider in their analyses (Arro, 2013; Rosenbaum & Valsiner, 2011). Here we have focused on behaviours, but it may well be possible that our observers also considered the macaques' outer appearances, such that some had more robust or more gracile bones and limbs or physiognomic peculiarities. The observers may have associated particular individual-specific outer appearances and behaviours with one another, such as assuming that physically more robust macaques may also be the more assertive ones. We do not know about this. Neither do we know how the observers actually interpreted the macaques' behaviours (and outer appearances), what explanations they may have developed for these, which inferences they may have drawn about possible causal events in both the macaques' psyche and the environmental situations that they encountered, and what appraisals the

observers may have made for all of this. We can only speculate about this because the fixed contents of personality inventories precluded the collection of any further information.

Alternative methods that are designed to match the theoretical nature and the complexity of the phenomena under study are therefore required (Omi, 2012; Toomela, 2009, 2011; Weber, 1949; Westen, 1996). Specifically, to explore personal and social representations, we need methods that allow the persons under study to explicate what entities they conceive and how they interpret, appraise, and explain the phenomena that they represent in these entities as is enabled by open answer formats (Kelly, 1955; for an empirical example see e.g., Park, 1986). To explore what people actually perceive, what they consider salient and why, and how they abstract from perceived events and develop representations of an individual's personality we also need methods that allow for investigations of psychological *processes*, rather than just of their *outcomes* (Gillespie & Zittaun, 2010; Jovchelovitch, 2007; Komatsu, 2012; Pillai, 2012; Westen, 1996). Because psychological phenomena are bound to the present (see Section 1.), this requires investigations that are as temporally close as possible to these processes as can be done, for example, with microgenetic methods (e.g., Wagoner, 2009; Valsiner, 1998, 2012). Such methods are also needed to explore what fields of meaning particular persons construct for particular standardised items in personality inventories, how they do this, and how they subjectively synthesise these fields to provide a numerical estimate as this has already been demonstrated for NEO-PI items (Rosenbaum & Valsiner, 2011; cf. also Michell, 2000, 2003; Schwarz, 1999; Wagoner & Valsiner, 2005).

The restricted answer formats of personality inventories produce comparably well-structured data. The internal consistencies of the four rating factors (as indicated by the factor loadings) were much higher than those of the behavioural composite measures of BR_xES-Approach-generated working constructs. This was true although the rating factors referred to a much greater diversity of behaviours than the working constructs, which summarised only functionally related behaviours. The taxonomic structures of the rating data were clearer and simpler than those of the behavioural data that yielded many further associations between working constructs not reflected in the observer judgements. This higher structuredness of rating data is not astonishing. It essentially reflects the patterns of logic that our human minds follow—both in how people perceive, (re)construct, represent, and judge individuals and in how researchers develop personality inventories by selecting only those rating items of meaningfully related content that yield empirically consistent data structures (Uher, 2013).

4.3.2 Non-lexical construct generation: The Behavioural Repertoire x Environmental Situations Approach

The present analyses on coherence of the observers' representations with the macaques' observable individual behaviours and on attribution biases were enabled by a new taxonomic approach, the BR_xES-Approach. Its non-lexical selection rationale allows researchers to systematically generate constructs of individual-specificity for comprehensive taxonomic investigations both of individual-specific behaviours and of people's pertinent representations. Importantly, it allows for generating constructs of individual-specific behaviours *independent* of the pertinent representations that particular socio-cultural and language communities or particular researchers have developed. This is not possible for lexical approaches because their rationale is based on the selection of lexical encodings of social representations that particular language communities have developed (cf. Goldberg, 1981; John et al., 1988). Lexical approaches therefore cannot disentangle the phenomena that are being represented from their socio-cultural interpretation, appraisal, and explanation.

The BR×ES-Approach starts taxonomic research from behavioural repertoires rather than from lexical repertoires to generate constructs of individual-specificity, and defines these constructs based on the behaviours' scientifically established functional relatedness rather than on lay people's (or researchers') pertinent representations reflecting possible interrelations and socio-culturally shared valences of behaviours. This opens up

unprecedented opportunities to systematically explore how individual-specific behaviours actually vary within and between particular (human and nonhuman) populations. Further taxonomic approaches are needed for investigations of individual-specific outer appearances to study how these are associated with the behavioural structures in particular populations and to explore how these phenomena and their (possible) associations are represented and lexically encoded by particular human communities. This also allows for systematic analyses of the ways in which people's representations and judgements of personality reflect biases that are derived from their socio-cultural stereotypes about relations between individuals' behaviours, appearances, and demographic status, such as ethnic group, age, gender/sex, or social position (Uher et al., 2013). Once established, such comprehensive taxonomic models can also guide research on causally related phenomena internal and external to the individual, and on the complex relations among all these different kinds of phenomena. Such guidance is important because internal phenomena are not directly perceivable and the enormous diversity of external phenomena complicates the identification of those that are causally related (for details, see Uher, 2013)

4.4 Summary and future directions

This research successfully demonstrated the application of a new philosophy-of-science paradigm for personality psychology that explicitly differentiates people's representations and their lexical encodings from the phenomena that are being represented and encoded—individual-specific behaviours and outer appearances, causally related phenomena internal (especially psychological) and external (environmental) to the individual, and (socio-culturally shared) interpretations, appraisals, and explanations of these phenomena.

Here we provided strong empirical evidence that crab-eating macaques exhibit a broad range of individual-specific behaviours that are notably stable across 12 and 24 months; these findings provide a solid basis for future explorations of their internal and external causes and consequences. We showed that limited observational experiences were sufficient for novices to quickly develop personality impressions of macaque individuals comparable to those of experts who had already known these macaques for some time. We discussed possible pathways in the formation of these representations and highlighted the likely impact of the observers' socially shared knowledge of *human* personality, which present-day humans acquire together with their native language. This knowledge is also reflected in the notable 12- and 24-month stabilities of the representations that *different* experts and *different* novices had developed in the three Study years. The systematic coherence between the macaques' individual-specific behaviours and the observers' pertinent representations found for several constructs argue that this knowledge of human personality is viable to some extent to allow people to develop accurate representations also of nonhuman individuals—at least of some species that are phylogenetically related, yet not necessarily endemic. But we also found profound attribution biases in the observers' personality judgements of these nonhuman individuals that likely reflect their socio-culturally shared stereotypes about age, sex/gender, and social position of human individuals.

Our results extend previous evidence that the Behavioural Repertoire $_x$ Environmental Situations Approach (Uher, 2008a, 2008b, 2011a, 2011b) constitutes a viable non-lexical approach for comprehensive taxonomic investigations of individual-specific behaviours independent of the pertinent representations that people (both lay people and researchers) have developed of them. Moreover, it also allows for systematic taxonomic investigations of these representations. Thus, the two kinds of phenomena can be investigated using the same methodological approach. Together with further taxonomic approaches for investigations of individual-specific outer appearances, and with systematic explorations of causally related phenomena, this opens up unprecedented possibilities to explore how persons from different socio-cultural and language communities perceive individuals, what they consider salient, and how they construct and represent individual-specificity and why. It will therefore be insightful to study various species that differ in phylogenetic relatedness and

similarity to humans. But it will ultimately be most insightful to study the diversity within humankind (Uher, 2013).

The new philosophy-of-science paradigm also highlighted important methodological and methodical implications. In this study, we used judgements made on predetermined assessment scales with restricted answer formats and we analysed the thus-obtained rating data following established standards of contemporary personality psychology—knowing that the phenomena that such data reflect cannot have the metric properties that are frequently attributed to them and that are required for such statistics. We are fully aware that the analyses presented may not adequately reflect the representations that our observers had actually developed of the macaque individuals—as this is true for any questionnaire-based study. Yet it was our aim to express critical concerns about these methods by demonstrating by their very application essential divergences from behavioural data and important limitations in their abilities to capture representations, to unravel their possible development, and to reveal how people actually generate quantitative personality judgements in a given moment. This would not have been possible had we used other methods because any divergences and limitations could then be attributed to these other methods.

It is certainly worthwhile to study larger samples of both crab-eating macaques and of human observers, as is true for any research. But it is far more important to study and to better understand the *processes* that underlie the remarkable human abilities to quickly recognise and to mentally and socially represent individual-specificity in behaviour (and outer appearances) in individual members of our own species and of some others. A better knowledge of these abilities—enabled by a new paradigm offering a comprehensive philosophy of science and new methodologies for personality psychology and by the systematic further development of alternative non-questionnaire-based methods—will advance our understanding of some of the most important socio-cognitive developments in the recent evolution of our species and of the ways in which we are in fact unique.

***

Glossary of terms from the new philosophy-of-science paradigm for personality psychology relevant to the present analyses

Behavioural composite construct measures: Composite variables reflecting differential scores of individual behavioural probabilities on the level of the BR_xES-Approach-generated *working constructs*. They were derived from aggregating all those *behavioural measurements* that were assigned to a particular construct based on the scientifically established functional relatedness of the behaviours. Composite measures for which temporal reliability between the two Study waves of a given Study year can be evidenced reflect summary scores of *individual-specific patterns of behaviour*.

Behavioural measurements: Measurement variables reflecting differential scores of individual probabilities to display specific behaviours in the social situations that we observed in the macaque groups. They were derived from aggregating the raw measurements across all five observation days within each Study wave to obtain time-relative probabilities for each individual. These aggregate scores were then *z*-standardised across all individuals within each Study wave to obtain scores reflecting differential scores in these behavioural probabilities. Temporal reliability between Study waves within a given Study year provides evidence that the behavioural measurements reflect *individual-specific patterns of behaviour*.

Individual-specific patterns of behaviour: Because behaviours are highly fluctuating, individuals can be characterised only in behavioural probabilities that differ between individuals in relatively stable ways. To reflect individual-specificity, differential patterns of behavioural probabilities must be stable across time periods longer than those in which the probabilities were first ascertained and in ways considered to be meaningful (e.g., defined via test-retest correlation scores between Study waves). Accordingly, *behavioural measurements* reflecting individual-specificity were derived from *1)* aggregation over repeated measurement occasions; *2)* standardisation of these aggregate scores; and *3)* evidence of temporal reliability in the standardised aggregate scores.

Mean rating scores: The *raw rating scores* were aggregated on different levels: *1)* on the level of *each Study wave,* scores were averaged across raters separately for the different rater groups (i.e., mean expert ratings, mean novice ratings) and from these we computed mean scores of both rater groups; *2)* on the level of *each Study year,* the mean scores of both rater groups were aggregated across Study waves; *3)* the mean scores per Study year of the behaviour-descriptive verb items were aggregated on the level of the BR_xES-Approach generated working constructs.

Non-lexical emic personality constructs: Constructs of individual-specificity (called *working constructs*) that are not derived from within the human repertoires of everyday language terms as in lexical research, but from within the scientifically described behavioural-ecological system of the population under study using the Behavioural Repertoire x Environmental Situations-Approach (BR_xES-Approach; see Section 2.3). We used BR_xES-Approach-generated constructs to study the *individual-specific behaviours* that they describe for the macaques under study; for this purpose, we operationalised them with *behavioural measurements*. We also used these constructs to study the representations that human observers developed of the macaques' individual-specific behaviours. Each construct was therefore operationalised in rating items (i.e., in sentences with trait-adjectives, TA, and behaviour-descriptive verbs, BV) that describe their contents in the everyday language of the raters (experts and novices) under study.

Rating factor scores: Scores for each macaque individual for each Study year on each of the four rating factors. These rating factors were statistically derived from exploratory R-factor analyses on all single rating items (i.e., trait-adjective items and

behaviour-descriptive verb items) using the mean combined scores of both rater groups.

Raw rating scores: The rating scores that the single raters provided for each macaque individual on each rating item on a 5-point frequency scale.

Working constructs: Constructs generated with the BR_xES-Approach (see *Non-lexical emic personality constructs*) based on the hypothetical assumption that *individual-specific patterns* can be found in behaviours of particular categories displayed in situations of particular categories. This hypothetical assumption must be empirically substantiated for each construct; otherwise the particular working construct has to be discarded. The *individual-specific behaviours* described in the working constructs and the pertinent representations that human observers develop of them can be further explored for their underlying structures and stabilities.

Appendix

References reviewed for behavioural and situational categories in the application of the Behavioural Repertoire x Environmental Situations Approach for crab-eating macaques (*Macaca fascicularis*)

Aldrich-Blake, F. P. G. (1980). Long-tailed macaques. In D. J. Chivers (Ed.), *Malayan Forest Primates* (pp. 147–165). New York, NY: Plenum.

Angst, W. (1974). *Das Ausdrucksverhalten des Javaneraffen Macaca fascicularis Raffles. Eine Einführung*. Berlin and Hamburg: Paul Parey.

Angst, W. (1975). Basic data and concepts on the social organization of *Macaca fascicularis*. In L. A. Rosenblum (Ed.), *Primate Behavior:Developments in Field and Laboratory Research, Vol. 4* (pp. 325–388). New York, NY: Academic Publishing.

Aureli, F. (1992). Post-conflict behaviour among wild long-tailed macaques (*Macaca fascicularis*). *Behavioral Ecology and Sociobiology, 31,* 329–337.

Aureli, F., Das, M. & Veenema, H. C. (1997). Differential kinship effect on reconciliation in three species of macaques (*Macaca fascicularis, M. fuscata, and M. sylvanus*). *Journal of Comparative Psychology, 111*, 91-99.

Cords, M. (1992). Post-conflict reunions and reconciliation in long-tailed macaques. *Animal Behaviour, 44,* 57–61.

Das, M., Penke, Z., & van Hooff, J. A. R. A. M. (1997). Affiliation between aggressors and third parties following conflicts in long-tailed macaques (*Macaca fascicularis*). *International Journal of Primatology*, 18, 157-179.

De Ruiter, J. & Geffen, E. (1998). Relatedness of matrilines, dispersing males and social groups in long-tailed macaques (*Macaca fascicularis*). *Proceedings of the Royal Society Biological Sciences, 285,* 79–87.

De Ruiter, J., van Hooff, J. & Scheffrahn, W. (1995). Social and genetic aspects of paternity in wild long-tailed macaques (*Macaca fascicularis*). *Behaviour, 129,* 203–224.

Dittus, W. (2004). Demography: A window to social evolution. In: B. Thierry, M. Singh, & W. Kaumanns (Eds.), *Macaque societies: A model for the study of social organization* (pp. 87-112). Cambridge, UK: Cambridge University Press.

Fittinghoff, N. A., & Lindburg, D. G. (1980). Riverine refuging, in East Bornean *Macaca fascicularis*. In D. G. Lindburg, (Ed.), *The macaques: Studies in ecology, behavior, and evolution*, (pp. 182–214). New York, NY: Van Nostrand Reinhold.

Fooden, J. (1995). Systematic review of southeast Asian longtail macaques, *macaca fascicularis* (Raffles, [1821]). *Fieldiana, Zoology New Series, 81,* 1-206.

Palombit, R. A. (1992). A preliminary study of vocal communication in wild long-tailed macaques (*Macaca fascicularis*). I. Vocal repertoire and call emission. *International Journal of Primatology, 13,* 143-82.

Palombit, R. A. (1992). A preliminary study of vocal communication in wild long-tailed macaques (*Macaca fascicularis*). II. Potential of calls to regulate intragroup spacing. *International Journal of Primatology, 13,* 183–207.

Poirier, F. E., & Smith, E. O. (1974). The crab-eating macaques (*Macaca fascicularis*) of Angaur Island, Palau, Micronesia. *Folia Primatologica, 22*, 258-306.

Tobin, H., Logue, A. W., Chelonis, J. J., Ackerman, K. T., & May III, J. G. (1996). *Self-control* in the monkey *Macaca fascicularis*. *Animal Learning and Behavior, 24,* 168-174.

van Noordwijk, M. A. (1985). Sexual behaviour of Sumatran long-tailed macaques (*Macaca fascicularis*). *Zeitschrift für Tierpsychologie, 70*, 277–296.

van Noordwijk, M. A., & van Schaik, C. P. (1987). Competition among female long-tailed macaques (*Macaca fascicularis). Animal Behaviour, 35,* 577–589.

van Schaik, C. P., van Amerongen, A. & van Noordwijk, M. A. (1996) Riverine refuging by wild Sumatran long-tailed macaques (*Macaca fascicularis*). In W. C. McGrew, T. Nishida, L. F. Marchant, J. E. Fa & D. G. Lindburg (Eds.), *Evolution and Ecology of Macaque Societies* (pp. 160-182). Cambridge, MA: Cambridge University Press.

van Schaik, C. P., van Noordwijk, M. A., de Boer, R. J., & den Tonkelaar, I. (1983). The effect of group size on time budgets and social behavior in wild long-tailed macaques. *Behavioral Ecology and Sociobiology, 13,* 173–181.

Veenema, H. C., Spruijt, B. M., Gispen, W. H., & van Hooff, J. A. R. A. M. (1997). Aging, dominance history, and social behavior in Java monkeys (*Macaca fascicularis*). *Neurobiology of Aging, 18,* 509-515.

Wheatley, B. P. (1980). Feeding and ranging of east Bornean *Macaca fascicularis*. In D. G. Lindburg (Ed.), *The macaques: Studies in ecology, behavior, and evolution* (pp. 215-246). New York, NY: Van Nostrand Reinhold.

Yeager, C. P. (1996). Feeding ecology of the long-tailed macaque (*Macaca fascicularis*) in Kalimantan Tengah, Indonesia. *International Journal of Primatology, 17*, 51-62.

References

Altmann, J. (1974). Observational study of behavior: Sampling methods. *Behaviour*, 49, 229–267.

Angst, W. (1974). *Das Ausdrucksverhalten des Javaneraffen Macaca fascicularis Raffles. Eine Einführung*. Berlin und Hamburg: Paul Parey.

Arro, G. (2013). Peeking into personality test answers: Inter- and intraindividual variety in item interpretations. *Integrative Psychological and Behavioral Science, 47,* 56-76.

Asch, S. E. (1946). Forming impressions of personality. *The Journal of Abnormal and Social Psychology, 41,* 258–290.

Asendorpf, J. B. (1988). Individual response profiles in the behavioral assessment of personality. *European Journal of Personality, 2,* 155-167.

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51*, 1173-1182.

Belyaev, D. K. (1969). Domestication of animals. *Science Journal (U.K.), 5*, 47–52.

Biesanz, J. C., & Human, L. J. (2010). The cost of forming more accurate impressions: Accuracy motivated perceivers see the personality of others more distinctively but less normatively. *Psychological Science, 24,* 589-594.

Blaffer-Hrdy, S. (2009). *Mothers and others: the evolutionary origins of mutual understanding*. Cambridge MA: Belknap Press.

Block, J. (2010). The five-factor framing of personality and beyond: Some ruminations. *Psychological Inquiry, 21,* 2–25.

Bock, P. K. (2000). Culture and personality revisited. *American Behavioural Scientist, 44*, 32-40.

Brown, D. E. (1991). *Human universals*. New York, NY: McGraw-Hill.

Capitanio, J. P. (1999). Personality dimensions in adult male rhesus macaques: Prediction of behaviors across time and situation. *American Journal of Primatology, 47,* 299-320.

Cervone, D., Shadel, W. G., & Jencius, S. (2001). *Social-cognitive theory of personality* assessment. *Personality and Social Psychology Review, 5,* 33-51.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112,* 155-159.

Daston, L. & Galison, P. (2007). *Objectivity.* New York, NY: Zone Books.

Diamond, J. (2002). Evolution, consequences and future of plant and animal domestication. *Nature, 418,* 700–707.

Diriwächter, R., Valsiner, J. & Sauck, C. (2004). Microgenesis in making sense of oneself: Constructive recycling of personality inventory items. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research, 6*, Art. 11.

Dunbar, R. (1996). *Grooming, gossip, and the evolution of language*. Cambridge, MA: Harvard University Press.

Furr, R. M. (2009). Personality psychology as a truly behavioural science. *European Journal of Personality, 23,* 369-401.

Germonpré, M., Sablin, M. V., Stevens, R. E., Hedges, R. E. M, Hofreiter, M., Stiller, M., & Després, V. R. (2009). Fossil dogs and wolves from Palaeolithic sites in Belgium, the Ukraine and Russia: Osteometry, ancient DNA and stable isotopes. *Journal of Archaeological Science, 36,* 473-490.

Gillespie, A. & Zittaun, T. (2010). Studying the moment of thought. In A. Toomela & J. Valsiner (Eds.), *Methodological thinking in psychology: 60 years gone astray*? (pp. 69-88). Charlotte, NC: Information Age Publishing.

Goldberg, L. R. (1981). Language and individual differences: The search for universals in personality lexicons. In L. Wheeler (Ed.), *Review of Personality and Social Psychology: Vol. 2* (pp. 141-165). Beverly Hills, CA: Sage.

Grice, J. W. (2001a). A comparison of factor scores under conditions of factor obliquity. *Psychological Methods, 6,* 67-83.

Grice, J. W. (2001b). Computing and evaluating factor scores. *Psychological Methods, 6,* 430-450.

Guidelines for the Treatment of Animals in Behavioral Research and Teaching (2006). Issued by Elsevier and incorporated by reference into the Animal Behavior Guide for Authors. *Animal Behavior, 71,* 245-253.

Gumert, M. D., Kluck, M., & Malaivijitnond, S. (2009). The physical characteristics and usage patterns of stone axe and pounding hammers used by long-tailed macaques in the Andaman Sea region of Thailand. *American Journal of Primatology,* 71**,** 594–608.

Heine*,* S. J., Lehman*,* D. R., Peng, K., & Greenholtz, J. (*2002*). What's wrong with cross-cultural comparisons of subjective Likert scales?: The reference-group effect. *Journal of Personality and Social Psychology, 82,* 903-918.

Hume, D. (1757/1957). *The natural history of religion*. Stanford, CA: Stanford University Press.

John, O. P., Angleitner, A., & Ostendorf, F. (1988). The lexical approach to personality: A historical review of trait taxonomic research. *European Journal of Personality, 2,* 171-203.

Jovchelovitch, S. (2007*). Knowledge in context: Representations, community and culture*. London, UK: Routledge.

Kelly, G. A. (1955). *The Psychology of Personal Constructs* (Vol. 1 and 2). New York, NY: Norton.

King, J. E., Weiss, A., & Sisco, M. S. (2008). Aping humans: Age and sex effects in chimpanzee (Pan troglodytes) and human (Homo sapiens) personality. *Journal of Comparative Psychology, 122,* 418-27.

Komatsu, K. (2012). *Temporal reticence of the self: who can know my self*? *Integrative Psychological and Behavioral Science*, 46, 357-372.

Lahlou, S. (1996). Propagation of social representations. *Journal for the Theory of Social Behaviour, 26,* 157-175.

Lahlou, S. (2001). Functional aspects of social representations. In K. Deaux & G. Philogene (Eds.), *Representations of the social: Bridging theoretical traditions,* (pp. 131-146). Oxford, UK: Blackwell.

Lahlou, S. (2008). *L'Installation du Monde: De la représentation à l'activité en situation*. Aix-en-Provence, Université de Provence: Habilitation à Diriger des Recherches en Psychologie, 375.

Lamiell, J. T. (2003). *Beyond individual and group differences: Human individuality, scientific psychology, and William Stern's critical personalism.* Thousand Oaks, California: Sage Publications.

Lewin, K. (1935). *A dynamic theory of personality. Selected papers*. New York, NY: McGraw-Hill.

Locke, J., (1689/1975). *Essay concerning human understanding. Book II*. (chapter 32). Oxford, UK: Oxford University Press.

MacDonald, D. W. (2001). *The New Encyclopedia of Mammals*. Oxford, UK: Oxford University Press.

McAdams, D. P. (1994). A psychology of the stranger. *Psychological Inquiry, 5*, 145-148.

Michell, J. (1997). Quantitative science and the definition of measurement in psychology. *British Journal of Psychology*, *88*, 355–383.

Michell, J. (2000). Normal science, pathological science and psychometrics. *Theory & Psychology*, *10*, 639–667.

Michell, J. (2003). The quantitative imperative: Positivism, naïve realism and the place of quantitative methods in psychology, *Theory & Psychology*, *13*, 5–31.

Millikan, R. (1993). *White queen psychology and other essays for Alice*. Bradford, MA: MIT Press.

Mischel, W. (1968). *Personality and assessment.* New York, NY: Wiley.

Mischel, W. (1977). The interaction of person and situation. In: D. Magnusson, & N. S. Endler (Eds.), *Personality at the crossroads: Current issues in interactional psychology* (pp. 333-352). Hillsdale, NJ: Erlbaum.

Mischel, W., & Shoda, Y. (1994). Personality psychology has two goals: Must it be two fields? *Psychological Inquiry, 5*, 156-158.

Moscovici, S. (1984) The phenomenon of social representations. In R. M. Farr & S. Moscovici (Eds.), *Social representations* (pp. 3-70). Cambridge, MA: Cambridge University Press.

Neuman, Y., Turney, P. D., & Cohen, Y. (2012), How language enables abstraction: A study in computational cultural psychology. *Integrative Psychological and Behavioral Science, 46,* 129-145.

Omi,Y. (2012). Tension between the theoretical thinking and the empirical method: Is it an inevitable fate for psychology? *Integrative Psychological and Behavioral Science, 46*, 118-127.

Park, B. (1986). A method for studying the development of impressions of real people. *Journal of Personality and Social Psychology, 51,* 907-917.

Pederson, A. K., King, J. E., & Landau, V. I. (2005). Chimpanzee (*Pan troglodytes*) personality predicts behavior. *Journal of Research in Personality, 39,* 534-549.

Peirce, C. S. (1902/1958). The simplest mathematics (CP 4.227-323). In *Collected papers of Charles Sanders Peirce (1931-1935),* Vol. 4, C. Hartshorne & P. Weiss (eds.), Cambridge, MA: Harvard University Press.

Pillai, P. (2012). Cultural directions and origins of everyday decisions. *Integrative Psychological and Behavioral Science, 46,* 235-242.

Rosenbaum, P. J., & Valsiner, J. (2011). The un-making of a method: From rating scales to the study of psychological processes. *Theory & Psychology, 21,* 47-65.

Schwarz, N. (1999). Self-reports: How the questions shape the answer. *American Psychologist, 54,* 93–105.

Sherman, J. W. & Klein, S. B. (1994). The development and representation of personality impressions. *Journal of Personality and Social Psychology, 67,* 972-983.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86,* 420-428.

Stern, W. (1911). *Die Differentielle Psychologie in ihren methodischen Grundlagen [Differential Psychology in its methodological foundations]*. Leipzig: Barth.

Stevenson-Hinde, J., Stillwell-Barnes, R., & Zunz, M. (1980). Individual differences in young rhesus monkeys: consistency and change. *Primates, 21,* 498-509.

Thurstone, L. L. (1935). *The vectors of mind*. Chicago, IL: University of Chicago Press.

Toomela, A. (2008). *Variables in psychology: A critique of quantitative psychology. Integrative Psychological and Behavioral Science, 42,* 245-265.

Toomela, A. (2009). How methodology became a toolbox – and how it escapes from that box. In J. Valsiner, P. Molenaar, M. Lyra, and N. Chaudhary (Eds.), *Dynamic process methodology in the social and developmental sciences* (pp. 45 - 66). New York, NY: Springer

Toomela, A. (2011). Travel into a fairy land: A critique of modern qualitative and mixed methods psychologies. *Integrative Psychological and Behavioral Science, 45*, 21-47.

Trendler, G. (2009). Measurement theory, psychology and the revolution that cannot happen. *Theory and Psychology, 19,* 579-599.

Trut, L. N. (1999). Early canid domestication: The farm-fox experiment. *American Scientist, 87,* 160–169.

Uher, J. (in preparation-a). What is behaviour? And (when) is language behaviour? A meta-theoretical definition.

Uher, J. (in preparation-b). Meta-theoretical foundations of objectivity versus subjectivity in quantifications of behaviour and personality.

Uher, J. (submitted for publication). Methodological approaches to taxonomies of human personality differences: The Behavioural Repertoire x Environmental Situations Approach - A non-lexical alternative.

Uher, J. (2008a). Comparative personality research: Methodological approaches. *European Journal of Personality, 22,* 427-455.

Uher, J. (2008b). Three methodological core issues of comparative personality research. *European Journal of Personality, 22,* 475-496.

Uher, J. (2011a). Individual behavioral phenotypes: An integrative meta-theoretical framework. Why 'behavioral syndromes' are not analogues of 'personality'. *Developmental Psychobiology, 53*, 521–548.

Uher, J. (2011b). Personality in nonhuman primates: What can we learn from human personality psychology? In A. Weiss, J. King, & L. Murray (Eds.), *Personality and Temperament in Nonhuman Primates* (pp. 41-76). New York, NY: Springer.

Uher, J. (2013). Personality psychology: Lexical approaches, assessment methods, and trait concepts reveal only half of the story—Why it is time for a paradigm shift. *Integrative Psychological and Behavioral Science, 47,* 1-55. DOI: 10.1007/s12124-013-9230-6

Uher, J., & Asendorpf, J. B. (2008). Personality assessment in the Great Apes: Comparing ecologically valid behavior measures, behavior ratings, and adjective ratings. *Journal of Research in Personality, 42,* 821-838.

Uher, J., Addessi, E., & Visalberghi, E. (2013). Contextualised behavioural measurements of personality differences obtained in behavioural tests and social observations in adult capuchin monkeys (*Cebus apella*). *Journal of Research in Personality.*

Uher, J., Asendorpf, J. B., & Call, J. (2008). Personality in the behaviour of great apes: Temporal stability, cross-situational consistency and coherence in response. *Animal Behaviour, 75,* 99-112.

Valsiner, J. (1998). *The guided mind*. Cambridge, MA: Harvard University Press.

Valsiner, J. (2012). *A guided science: History of psychology in the mirror of its making*. New Brunswick, NJ: Transaction Publishers.

Veltkamp, G. M., Recio, G., & Jacobs, A. M., & Conrad, M. (2012). Is personality modulated by language? Evidence from bilinguals' NEO-FFI scores. *International Journal of Bilingualism.* Published online before print April 16 ,2012, doi:10.1177/1367006912438894.

Visalberghi, E., & Fragaszy, D. (1990). Food-Washing behaviour in tufted capuchin monkeys, *Cebus apella*, and crabeating macaques, *Macaca fascicularis*. *Animal Behaviour*, 40, 829-836.

Vygotsky, L. S. (1962). *Thought and language*. Cambridge, MA: MIT Press.

Wagner, W., Farr, R., Jovchelovitch, S., Lorenzi-Cioldi, F., Marková, I., Duveen, G., & Rose, D. (1999). Theory and method of social representations. *Asian Journal of Social Psychology, 2,* 95-125.

Wagoner, B. (2009). The experimental methodology of constructive microgenesis. In: J. Valsiner, P. Molenaar, N. Chaudhary, and M. Lyra (Eds.). *Handbook of dynamic process methodology in the social and developmental sciences* (pp. 99-121). New York, NY: Springer.

Wagoner, B., & Valsiner, J. (2005). Rating tasks in psychology: From static ontology to dialogical synthesis of meaning. In A. Gülerce, I. Hofmeister, G. Saunders, & J. Kaye (Eds.), *Contemporary theorizing in psychology: Global perspectives* (pp. 197–213). Toronto, Canada: Captus.

Weber, M. (1949). *The methodology of the social sciences* (E. Shils & H. Finch, Trans., Eds.). New York, NY: Free Press.

Westen, D. (1996). A model and a method for uncovering the nomothetic from the idiographic: An alternative to the Five-Factor Model. *Journal of Research in Personality, 30,* 400–413.

Wundt, W. M. (1904). *Principles of physiological psychology*. London, UK: Allen.

Tables

Table 1 Test-retest reliability of the behavioural composite measures on the level of BR_xES-Approach-generated working constructs

Working construct	t_1-t_2	t_3-t_4	t_5-t_6
Aggressiveness to group members	**.30** (.006)	**.45** (<.001)	.09 (.243)
Anxiousness	**.22** (.032)	**.29** (.013)	**.25** (.030)
Arousability	**.48** (<.001)	**.79** (<.001)	**.80** (<.001)
Cleanliness	**.51** (<.001)	**.76** (<.001)	**.55** (<.001)
Curiousness	**.89** (<.001)	**.75** (<.001)	**.79** (<.001)
Dominance	.10 (.215)	**.39** (.001)	.05 (.361)
Gregariousness	**.50** (<.001)	**.77** (<.001)	**.93** (<.001)
Impulsiveness	**.24** (.025)	**.63** (<.001)	.13 (.232)
Playfulness	**.71** (<.001)	**.87** (<.001)	**.87** (<.001)
Sexual activity	**.75** (<.001)	**.50** (<.001)	**.73** (<.001)
Social orientation to group members	**.68** (<.001)	**.40** (.001)	**.68** (<.001)
Third-party intervention	-	**-**	.23 (.058)
Mean	.49	.60	.51

Note. Pearson correlations *r* based on $n = 70$ individuals in t_1-t_2; $n = 60$ individuals in t_3-t_4; and $n = 34$-58 individuals in t_5-t_6 (in Study year 3, some measures were obtained in only one group). Significant correlations are bold; one-sided *p* values in parentheses.

Table 2 Average interrater and test-retest reliability of the observer judgements on trait-adjective items (TA) and behaviour-descriptive verb (BV) items of the Macaque Personality Inventory for captive populations (MPIc) across Study waves – Summative statistics

Item format	Interrater reliability													
	Experts				Novices						The two rater groups combined			
	t_1	t_2	t_3	t_6	t_1	t_2	t_3	t_4	t_5	t_6	t_1	t_2	t_3	t_6
	ICC(3,*k*)													
Trait-adjectives	.59	.59	.61	.60	.64	.66	.57	.69	.69	.65	.72	.75	.72	.73
Behaviour-descriptive verbs	.66	.52	.66	.63	.65	.64	.52	.60	.60	.63	.75	.73	.69	.70
All items	.63	.55	.64	.62	.65	.65	.54	.64	.64	.64	.74	.74	.70	.71
	ICC(3,1)													
Trait-adjectives	.43	.43	.46	.45	.50	.50	.42	.54	.54	.49	.41	.46	.41	.42
Behaviour-descriptive verbs	.50	.39	.51	.47	.49	.49	.36	.45	.45	.47	.45	.42	.37	.38
All items	.48	.40	.49	.46	.49	.49	.38	.49	.49	.48	.44	.43	.39	.40

	Test-retest reliability *r*									
	Experts			Novices			The two rater groups combined	Cross-correlations between the two rater groups			
	t_1-t_2			t_1-t_2	t_3-t_4	t_5-t_6	t_1-t_2	Experts t_1-Novices t_2	Novices t_1-Experts t_2	Experts t_3-Novices t_4	Novices t_5-Experts t_6
Trait-adjectives	.76	-	-	.53	.49	.54	.75	.56	.51	.62	.59
Behaviour-descriptive verbs	.79	-	-	.47	.37	.52	.73	.59	.51	.65	.58
All items	.78	-	-	.49	.42	.52	.74	.54	.52	.60	.59

Note. Interrater reliability: The *ICC*(3,*k*) scores for expert ratings and novice ratings are based on $k = 2$ raters in each Study wave; those of the combined ratings of both rater groups are based on $k = 4$ raters. Across all Study waves, scores for expert ratings are based on $n = 34$ to 97 macaque individuals, those of novice ratings and of the combined ratings of both rater groups are based on $n = 34$ to 49 macaque individuals (see Section 2.6). Test-retest reliability: Single item correlations for expert ratings and combined ratings are based on $n = 45$, and for novice ratings on $n = 40$ in t_1-t_2; on $n = 24$ in t_3-t_4; and on $n = 49$ in t_5-t_6.

Table 3 Cross-method coherence and mediation analyses on the level of the BR$_x$ES-Approach-generated working constructs

Working constructs	Coherence between methods[a]						Mediation analyses between methods[b]			
	EO-BV		BV-TA		EO-TA		EO-*(BV)*-TA		BV-*(EO)*-TA	
Aggressiveness to group members	**.58**	(<.001)	**.90**	(<.001)	**.61**	(<.001)	**.15**	(<.001)	**.80**	(<.001)
Arousability	**.31**	(.003)	**.16**	(.050)	**.20**	(.038)	**.25**	(.039)	-.15	(.215)
Anxiousness	**.54**	(<.001)	**.73**	(<.001)	**.48**	(<.001)	.10	(.246)	**.70**	(<.001)
Cleanliness	**.46**	(<.001)	**.73**	(<.001)	**.54**	(<.001)	**.24**	(.003)	**.65**	(<.001)
Competitiveness	--	--	**.78**	(<.001)	--	--	--	--	--	--
Curiousness	**.26**	(.011)	**.87**	(<.001)	**.23**	(.023)	.00	(.952)	**.86**	(<.001)
Distractibility	--	--	**.76**	(<.001)	--	--	--	--		
Dominance	**.52**	(<.001)	**.81**	(<.001)	**.41**	(<.001)	.00	(.968)	**.80**	(<.001)
Gregariousness	**.61**	(<.001)	**.78**	(<.001)	**.51**	(<.001)	.02	(.779)	**.79**	(<.001)
Impulsiveness	**.42**	(<.001)	**.27**	(.003)	**.46**	(<.001)	**.40**	(.001)	**.13**	(.248)
Physical activity	--	--	**.93**	(<.001)	--	--	--	--		
Persistence	--	--	-.35	(1.00)	--	--	--	--		
Playfulness	**.77**	(<.001)	**.95**	(<.001)	**.82**	(<.001)	**.19**	(.008)	**.81**	(<.001)
Sexual activity	**.39**	(<.001)	**.79**	(<.001)	**.53**	(<.001)	**.23**	(<.001)	**.75**	(<.001)
Social orientation to group members	**.37**	(<.001)	**.28**	(.002)	.00	(.996)	-.12	(.296)	**.34**	(.006)
Social orientation to youngsters	--	--	**.55**	(<.001)	--	--	--	--	--	--
Third-party intervention	--	--	**.90**	(<.001)	--	--	--	--	--	--
Vigilance	--	--	**.78**	(<.001)	--	--	--	--	--	--

Note. Methods: EO – behavioural composite construct measures obtained in ethological observations, TA – trait-adjective ratings, BV – behaviour-descriptive verb ratings. Mediation analyses: EO-BV – Correlations of predictors (EO) with potential mediators (BV); EO-TA – Correlations of predictors (EO) with potential criteria (TA); EO-*(BV)*-TA – Regression coefficients of predictors (EO) on criteria (TA) controlled for mediators (BV); BV-*(EO)*-TA –regression coefficients of mediators (BV) on criteria (TA) controlled for predictors (EO). Correlations of trait-adjective ratings and behaviour-descriptive verb ratings based on $N = 104$, correlations with behavioural composite construct measures based on $n = 78$. Significant coefficients are bold; p values in parentheses (correlations one-sided, regression coefficients two-sided). [a] Pearson correlations r; [b] Standardised regression coefficients β in multiple regression equations.

Table 4 Intercorrelation matrices of behavioural composite measures of BR_xES-Approach-generated working constructs in Study years 1 and 2

Working constructs		PL	CU	IM	AR	AG	SX	GR	SO	AX	DO	SC
Playfulness	PL		.25 (.056)	**.36** (.004)	.07 (.589)	.06 (.621)	-.05 (.724)	-.08 (.560)	**-.31** (.017)	-.02 (.900)	**.55** (.000)	-.07 (.603)
Curiousness	CU	**.33** (.005)		**.49** (.000)	**.45** (.000)	**.28** (.027)	**.45** (.000)	-.18 (.171)	-.22 (.086)	.21 (.098)	.12 (.371)	.00 (.973)
Impulsiveness	IM	**.46** (.000)	**.48** (.000)		**.30** (.012)	**.43** (.000)	.08 (.516)	-.01 (.937)	-.06 (.669)	.23 (.076)	-.01 (.945)	-.10 (.424)
Arousability	AR	.15 (.212)	**.54** (.000)	**.26** (.044)		**.50** (.000)	**.38** (.002)	.00 (.974)	**-.30** (.020)	-.15 (.249)	.10 (.428)	.14 (.269)
Aggressiveness to group members	AG	.13 (.290)	.21 (.081)	**.41** (.000)	**.48** (.000)		**.38** (.003)	**.32** (.012)	.12 (.339)	.12 (.343)	-.16 (.222)	.02 (.905)
Sexual activity	SX	.18 (.130)	**.55** (.000)	**.38** (.001)	**.52** (.000)	**.49** (.000)		-.08 (.559)	-.06 (.667)	.10 (.464)	-.11 (.417)	-.08 (.530)
Social orientation to group members	SO	.12 (.325)	.13 (.267)	**.40** (.001)	**.48** (.000)	**.75** (.000)	**.48** (.000)		.21 (.077)	-.09 (.508)	-.15 (.264)	**.29** (.025)
Gregariousness	GR	-.16 (.175)	**-.26** (.027)	-.17 (.157)	-.20 (.101)	.00 (.990)	-.01 (.918)	**.39** (.002)		**.52** (.000)	**-.40** (.001)	**-.30** (.019)
Dominance	DO	-.18 (.141)	.07 (.582)	-.02 (.894)	-.19 (.119)	.01 (.910)	-.03 (.788)	-.21 (.083)	**.36** (.002)		**-.57** (.000)	**-.51** (.000)
Anxiousness	AX	**.43** (.000)	.17 (.153)	.17 (.151)	-.03 (.796)	-.18 (.141)	.02 (.876)	.01 (.940)	**-.25** (.039)	**-.36** (.004)		.12 (.375)
Cleanliness	SC	.07 (.543)	.16 (.184)	.22 (.063)	**.58** (.000)	**.50** (.000)	.17 (.160)	**.49** (.000)	**-.24** (.045)	**-.29** (.014)	**.25** (.036)	

Note. BR_xES-Approach-generated working constructs sorted by their interrelations as reflected in the rating factor scores. Below diagonal intercorrelations from Study year 1 based on n = 71 macaque individuals; above diagonal intercorrelations from Study year 2 based on n = 61 macaque individuals. Pearson correlations *r*; significant correlations ($p < .05$; two-sided) are bold. Highlighted cells indicate associations between working constructs that are reflected in the four-factorial rating structure.

Table 5 Exploratory factor analysis of observer judgements on trait-adjective ratings (TA) and behaviour-descriptive verb ratings (BV) using the Macaque Personality Inventory for captive populations (MPIc): Item contents, assignments to the BR_xES-Approach-generated working constructs, factor loadings, and communalities

Working construct	Item format[a]	Item content (abbreviated)	Rating factors				Item communality
			Playful-active-curious	Aggressive-competitive	Prosocial-gregarious	Assertive-nonanxious	
Playfulness	TA	Playful	**.91**	-.25	.10	.05	.91
	BV	Plays alone, also with objects	**.96**	-.29	-.12	.13	.86
	BV	Plays with his/her group members	**.89**	-.24	.07	.10	.85
Curiousness	TA	Curious	**.87**	-.06	.07	.14	.82
	BV	Inspects and touches new objects	**.90**	.02	-.12	.10	.75
	BV	Ignores new objects	**-.89**	.01	.06	-.17	.77
Physical activity	TA	Physically active	**.92**	.00	-.01	.00	.84
	BV	Walks or brachiates	**.93**	.02	-.11	-.02	.81
	BV	Sits or lays around	**-.85**	-.18	.02	.12	.78
Social orientation to youngsters	TA	Friendly to youngsters	**.41**	**-.42**	.22	.10	.42
	BV	Plays with youngsters	**.81**	-.26	.18	.05	.80
Distractibility	TA	Distractible	**.78**	-.06	-.02	-.29	.69
	BV	Lets him/herself easily interrupt in his/her activities	**.69**	-.19	-.04	-.33	.60
Vigilance	TA	Vigilant	**.55**	.32	.00	-.06	.45
	BV	Spots small changes quickly	**.69**	.25	.06	-.01	.62
Arousability	TA	Excitable	**.59**	**.44**	.10	**-.40**	.72
Impulsiveness	TA	Impulsive	**.82**	.23	.06	.03	.82
Aggressiveness to group members	TA	Aggressive	-.15	**.88**	.05	.04	.78
	BV	Threatens or chases others	-.09	**.88**	.13	.03	.79
	BV	Hits or bites others	-.06	**.84**	.15	-.01	.71
Competitiveness	TA	Competitive	.29	**.83**	.10	.01	.88
	BV	Tries to get others' food or social partners	**.44**	**.57**	.12	.18	.71
Third-party intervention	TA	Intervening	-.22	**.84**	-.03	.11	.77

	BV	Tries to avoid getting involved in others' conflicts	.00	**-.74**	-.21	-.23	.74
	BV	Supports others who are involved in a conflict	-.24	**.84**	.06	-.05	.70
	BV	Tries to settle disputes among others	-.10	**.60**	-.25	.05	.44
Sexual activity	TA	Sexually active	-.05	**.75**	-.16	-.20	.58
	BV	Tries to contact others sexually	.02	**.76**	-.23	-.12	.61
	BV	Stimulates him/herself sexually	.34	.32	-.30	-.21	.30
Dominance	TA	Dominant	-.01	**.76**	.08	.37	.86
Social orientation to group members	TA	Friendly to group members	.21	**-.50**	.18	.05	.32
Anxiousness	BV	Stays away from unknown objects or persons	.06	**-.76**	.12	-.38	.82
Impulsiveness	BV	Shakes trees, jumps on or slaps others all of a sudden	.09	**.74**	-.05	.23	.68
Persistency	TA	Persistent	.31	**.62**	-.05	.36	.73
Gregariousness	TA	Gregarious	.04	-.07	**.90**	-.09	.80
	BV	Spends much time alone	.01	-.20	**-.83**	-.05	.76
	BV	Sits next to others	-.01	.01	**.87**	.12	.83
Social orientation to group members	BV	Approaches others, touches and also grooms them	.02	.13	**.78**	-.13	.60
	BV	Has bodily contact with others	.00	.01	**.88**	.09	.83
Social orientation to youngsters	BV	Takes care of youngsters by grooming or embracing them	-.26	.30	**.48**	-.33	.27
Dominance	BV	Can select the best place	.10	**.47**	.11	**.61**	.80
	BV	Makes way for others, displays bared teeth	-.05	-.38	-.06	**-.70**	.78
Anxiousness	TA	Anxious	-.07	-.26	.03	**-.82**	.82
	BV	Screams quickly, urinates and has at times also diarrhoea in conflict situations	.01	.00	.01	**-.72**	.52

Note. Based on mean ratings aggregated across 4 raters from the two rater groups for $N = 104$ macaque individuals, principal axis factoring, and promax rotation. [a] Item format: TA - Trait-adjective item, BV - Behaviour-descriptive verb item. Factor loadings ≥ .40 in absolute value are bold.

Table 6 Correlations between behavioural composite measures of BR_xES-Approach-generated working constructs and rating factor scores

Behavioural composite measures of working constructs	Rating factor scores: Playful-active-curious		Aggressive-competitive		Prosocial-gregarious		Assertive-nonanxious	
Playfulness	**.81**	(<.001)	-.08	(.491)	**.49**	(<.001)	.16	(.172)
Arousability	.02	(.871)	**.29**	(.009)	-.05	(.637)	.20	(.083)
Curiousness	**.25**	(.024)	**.34**	(.002)	-.14	(.208)	.18	(.113)
Impulsiveness	**.45**	(<.001)	**.28**	(.012)	**.26**	(.023)	.14	(.225)
Aggressiveness to group members	.10	(.393)	**.63**	(<.001)	.13	(.265)	**.28**	(.013)
Sexual activity	.08	(.487)	**.51**	(<.001)	.02	(.857)	.20	(.079)
Gregariousness	.10	(.372)	.02	(.873)	**.57**	(<.001)	**.37**	(.001)
Social orientation to group members	.06	(.575)	**.36**	(.001)	**.27**	(.017)	.11	(.358)
Anxiousness	**.31**	(.006)	**-.33**	(.003)	-.12	(.277)	**-.51**	(<.001)
Dominance	.09	(.424)	.21	(.061)	**.33**	(.003)	**.63**	(<.001)
Cleanliness	-.04	(.714)	.05	(.656)	-.12	(.294)	-.06	(.621)

Note. BR_xES-Approach-generated working constructs sorted by their interrelations as reflected in the rating factor scores. Pearson correlations *r* based on $n = 78$ individuals. Significant correlations are bold, two-sided *p* values in parentheses.

Table 7 Associations of demographic factors with the macaques' individual-specific behaviours studied with ethological observations (EO) and their reflection in the observers' representations studied with behaviour-descriptive verb ratings (BV) and trait-adjective ratings (TA) on the level of BR_xES-Approach-generated working constructs

Working constructs	Age			Male			Social rank		
Measure	**EO**	**BV**	**TA**	**EO**	**BV**	**TA**	**EO**	**BV**	**TA**
Aggressiveness to group members	-.08(.490)	.04(.664)	.10(.245)	**-.14**(.020)	-.23(.100)	**-.25**(.004)	**.53**(.000)	**.72**(.000)	**.72**(.000)
Arousability	-.06(.614)	.02(.847)	**-.56**(.000)	.20(.104)	.14(.246)	.01(.924)	**.25**(.037)	**-.34**(.006)	**.45**(.000)
Anxiousness	**-.32**(.002)	.02(.777)	-.07(.413)	**.21**(.039)	.06(.438)	-.08(.346)	**-.44**(.000)	**-.79**(.000)	**-.70**(.000)
Cleanliness	-.05(.654)	.00(.976)	-.10(.434)	.10(.418)	**-.34**(.007)	-.21(.091)	.06(.624)	-.05(.692)	**.26**(.037)
Competitiveness	-	**-.42**(.000)	**-.30**(.001)	-	-.14(.144)	-.12(.168)	-	**.67**(.000)	**.75**(.000)
Curiousness	-.03(.726)	**-.64**(.000)	**-.57**(.000)	**.60**(.000)	**.39**(.000)	**.35**(.000)	**.20**(.043)	**.20**(.010)	**.23**(.000)
Distractibility	-	**-.50**(.000)	**-.60**(.000)	-	.10(.328)	.07(.510)	-	-.17(.112)	.02(.881)
Dominance	.18(.109)	-.04(.623)	-.02(.784)	.16(.152)	-.06(.452)	-.11(.112)	**.40**(.001)	**.81**(.000)	**.87**(.000)
Gregariousness	-.02(.890)	-.10(.375)	-.12(.335)	-.19(.126)	-.00(.972)	-.06(.605)	**.18**(.027)	**.44**(.000)	**.33**(.007)
Impulsiveness	-.14(.253)	-.03(.733)	**-.52**(.000)	**.30**(.013)	.03(.741)	.08(.408)	.15(.194)	**.74**(.000)	**.38**(.000)
Physical activity	-	**-.72**(.000)	**-.67**(.000)	-	**.19**(.020)	**.24**(.004)	-	**.22**(.010)	**.22**(.012)
Persistence	-	**.57**(.000)	**-.24**(.014)	-	.11(.317)	-.10(.280)	-	**-.27**(.011)	**.69**(.000)
Playfulness	**-.41**(.001)	**-.63**(.000)	**-.57**(.000)	**.26**(.020)	**.37**(.000)	**.39**(.000)	.05(.656)	-.01(.908)	-.03(.722)
Sexual activity	-.06(.640)	-.13(.212)	-.13(.270)	.03(.826)	**.29**(.009)	-.12(.314)	.24(.056)	**.43**(.000)	**.42**(.001)
Social orientation to group members	-.13(.286)	-.17(.150)	-.05(.672)	**-.24**(.042)	-.07(.578)	**.39**(.001)	**.36**(.003)	**.38**(.001)	-.15(.196)
Social orientation to youngsters	-	**-.37**(.002)	**-.35**(.001)	-	.19(.108)	**.48**(.000)	-	.18(.125)	-.12(.207)
Third-party intervention	-	-.09(.304)	-.04(.623)	-	**-.29**(.001)	**-.28**(.002)	-	**.75**(.000)	**.75**(.000)
Vigilance	-	**-.64**(.000)	**-.50**(.000)	-	.16(.054)	.09(.376)	-	**.43**(.000)	**.44**(.000)

Note. EO – Behavioural composite construct measures obtained in ethological observations; BV – Behaviour-descriptive verb ratings aggregated on the level of the BR_xES-Approach generated working constructs; TA – Trait-adjective ratings. Standardised regression coefficients β in multiple regression equations. Significant coefficients are bold; two-sided p values in parentheses. Data from Study year 1 based on $n = 69$.

Table 8 Stability of behavioural composite measures of BR$_x$ES-Approach-generated working constructs and of rating factor scores across 12 and 24 months

Type of construct measure	12-month stabilities				24-month stabilities	
	Study years 1→2		Study years 2→3		Study years 1→3	
Behavioural composite measures of working constructs						
Playfulness	**.86**	(<.001)	-.04	(.574)	-.04	(.577)
Arousability	**.35**	(.003)	**.46**	(.004)	**.46**	(.005)
Curiousness	**.75**	(<.001)	**.83**	(<.001)	**.66**	(<.001)
Impulsiveness	**.46**	(<.001)	**.50**	(.002)	**.75**	(<.001)
Aggressiveness to group members	**.60**	(<.001)	**.64**	(<.001)	**.67**	(<.001)
Sexual activity	**.47**	(<.001)	**.60**	(<.001)	**.85**	(<.001)
Gregariousness	**.51**	(<.001)	**.50**	(.002)	**.42**	(.011)
Social orientation to group members	**.50**	(<.001)	**.71**	(<.001)	**.76**	(<.001)
Anxiousness	**.63**	(<.001)	**.35**	(.025)	**.55**	(.001)
Dominance	**.31**	(.009)	**.68**	(<.001)	**.75**	(<.001)
Cleanliness	.19	(.073)	**.72**	(<.001)	**.49**	(.003)
Mean	.54		.57		.61	
Rating factor scores						
Playful-active-curious	**.93**	(<.001)	**.68**	(<.001)	**.73**	(<.001)
Aggressive-competitive	**.75**	(<.001)	**.59**	(<.001)	**.56**	(<.001)
Prosocial-gregarious	**.59**	(<.001)	**.60**	(<.001)	**.41**	(.001)
Assertive- nonanxious	**.65**	(<.001)	**.52**	(.002)	**.40**	(.002)
Mean	.77		.60		.54	

Note. BR$_x$ES-Approach-generated working constructs sorted by their interrelations as reflected in the rating factor scores. Pearson correlations *r* for stability analyses between Study years 1→2 are based on *n* = 59 individuals and between Study years 2→3 on *n* = 30-31 for both behavioural measures and rating factor scores. Stability correlations between Study years 1→3 are based on *n* = 30 for behavioural measures and on *n* = 53 for rating factor scores. Significant correlations are bold; one-sided *p* values in parentheses.